RESEARCH CENTER

FIELD

Activity Report 2019

Section Scientific Foundations

Edition: 2020-03-21

ALGORITHMICS, PROGRAMMING, SOFTWARE AND ARCHITECTURE

ANTIQUE Project-Team

3. Research Program

3.1. Semantics

Semantics plays a central role in verification since it always serves as a basis to express the properties of interest, that need to be verified, but also additional properties, required to prove the properties of interest, or which may make the design of static analysis easier.

For instance, if we aim for a static analysis that should prove the absence of runtime error in some class of programs, the concrete semantics should define properly what error states and non error states are, and how program executions step from a state to the next one. In the case of a language like C, this includes the behavior of floating point operations as defined in the IEEE 754 standard. When considering parallel programs, this includes a model of the scheduler, and a formalization of the memory model.

In addition to the properties that are required to express the proof of the property of interest, it may also be desirable that semantics describe program behaviors in a finer manner, so as to make static analyses easier to design. For instance, it is well known that, when a state property (such as the absence of runtime error) is valid, it can be established using only a state invariant (i.e., an invariant that ignores the order in which states are visited during program executions). Yet searching for trace invariants (i.e., that take into account some properties of program execution history) may make the static analysis significantly easier, as it will allow it to make finer case splits, directed by the history of program executions. To allow for such powerful static analyses, we often resort to a *non standard semantics*, which incorporates properties that would normally be left out of the concrete semantics.

3.2. Abstract interpretation and static analysis

Once a reference semantics has been fixed and a property of interest has been formalized, the definition of a static analysis requires the choice of an *abstraction*. The abstraction ties a set of *abstract predicates* to the concrete ones, which they denote. This relation is often expressed with a *concretization function* that maps each abstract element to the concrete property it stands for. Obviously, a well chosen abstraction should allow one to express the property of interest, as well as all the intermediate properties that are required in order to prove it (otherwise, the analysis would have no chance to achieve a successful verification). It should also lend itself to an efficient implementation, with efficient data-structures and algorithms for the representation and the manipulation of abstract predicates. A great number of abstractions have been proposed for all kinds of concrete data types, yet the search for new abstractions is a very important topic in static analysis, so as to target novel kinds of properties, to design more efficient or more precise static analyses.

Once an abstraction is chosen, a set of *sound abstract transformers* can be derived from the concrete semantics and that account for individual program steps, in the abstract level and without forgetting any concrete behavior. A static analysis follows as a result of this step by step approximation of the concrete semantics, when the abstract transformers are all computable. This process defines an *abstract interpretation* [22]. The case of loops requires a bit more work as the concrete semantics typically relies on a fixpoint that may not be computable in finitely many iterations. To achieve a terminating analysis we then use *widening operators* [22], which over-approximate the concrete union and ensure termination.

A static analysis defined that way always terminates and produces sound over-approximations of the programs behaviors. Yet, these results may not be precise enough for verification. This is where the art of static analysis design comes into play through, among others:

- the use of more precise, yet still efficient enough abstract domains;
- the combination of application-specific abstract domains;
- the careful choice of abstract transformers and widening operators.

3.3. Applications of the notion of abstraction in semantics

In the previous subsections, we sketched the steps in the design of a static analyzer to infer some family of properties, which should be implementable, and efficient enough to succeed in verifying non trivial systems.

The same principles can be applied successfully to other goals. In particular, the abstract interpretation framework should be viewed as a very general tool to *compare different semantics*, not necessarily with the goal of deriving a static analyzer. Such comparisons may be used in order to prove two semantics equivalent (i.e., one is an abstraction of the other and vice versa), or that a first semantics is strictly more expressive than another one (i.e., the latter can be viewed an abstraction of the former, where the abstraction actually makes some information redundant, which cannot be recovered). A classical example of such comparison is the classification of semantics of transition systems [21], which provides a better understanding of program semantics in general. For instance, this approach can be applied to get a better understanding of the semantics of a programming language, but also to select which concrete semantics should be used as a foundation for a static analysis, or to prove the correctness of a program transformation, compilation or optimization.

3.4. From properties to explanations

In many application domains, we can go beyond the proof that a program satisfies its specification. Abstractions can also offer new perspectives to understand how complex behaviors of programs emerge from simpler computation steps. Abstractions can be used to find compact and readable representations of sets of traces, causal relations, and even proofs. For instance, abstractions may decipher how the collective behaviors of agents emerge from the orchestration of their individual ones in distributed systems (such as consensus protocols, models of signaling pathways). Another application is the assistance for the diagnostic of alarms of a static analyzer.

Complex systems and software have often times intricate behaviors, leading to executions that are hard to understand for programmers and also difficult to reason about with static analyzers. Shared memory and distributed systems are notorious for being hard to reason about due to the interleaving of actions performed by different processes and the non-determinism of the network that might lose, corrupt, or duplicate messages. Reduction theorems, e.g., Lipton's theorem, have been proposed to facilitate reasoning about concurrency, typically transforming a system into one with a coarse-grained semantics that usually increases the atomic sections. We investigate reduction theorems for distributed systems and ways to compute the coarse-grained counter part of a system automatically. Compared with shared memory concurrency, automated methods to reason about distributed systems have been less investigated in the literature. We take a programming language approach based on high-level programming abstractions. We focus on partially-synchronous communication closed round-based models, introduced in the distributed algorithms community for its simpler proof arguments. The high-level language is compiled into a low-level (asynchronous) programming language. Conversely, systems defined under asynchronous programming paradigms are decompiled into the high-level programming abstractions. The correctness of the compilation/decompilation process is based on reduction theorems (in the spirit of Lipton and Elrad-Francez) that preserve safety and liveness properties.

In models of signaling pathways, collective behavior emerges from competition for common resources, separation of scales (time/concentration), non linear feedback loops, which are all consequences of mechanistic interactions between individual bio-molecules (e.g., proteins). While more and more details about mechanistic interactions are available in the literature, understanding the behavior of these models at the system level is far from easy. Causal analysis helps explaining how specific events of interest may occur. Model reduction techniques combine methods from different domains such as the analysis of information flow used in communication protocols, and tropicalization methods that comes from physics. The result is lower dimension systems that preserve the behavior of the initial system while focusing of the elements from which emerges the collective behavior of the system.

The abstraction of causal traces offer nice representation of scenarios that lead to expected or unexpected events. This is useful to understand the necessary steps in potential scenarios in signaling pathways; this is useful as well to understand the different steps of an intrusion in a protocol. Lastly, traces of computation of

a static analyzer can themselves be abstracted, which provides assistance to classify true and false alarms. Abstracted traces are symbolic and compact representations of sets of counter-examples to the specification of a system which help one to either understand the origin of bugs, or to find that some information has been lost in the abstraction leading to false alarms.

ARIC Project-Team

3. Research Program

3.1. Efficient and certified approximation methods

3.1.1. Safe numerical approximations

The last twenty years have seen the advent of computer-aided proofs in mathematics and this trend is getting more and more important. They request: fast and stable numerical computations; numerical results with a guarantee on the error; formal proofs of these computations or computations with a proof assistant. One of our main long-term objectives is to develop a platform where one can study a computational problem on all (or any) of these three levels of rigor. At this stage, most of the necessary routines are not easily available (or do not even exist) and one needs to develop *ad hoc* tools to complete the proof. We plan to provide more and more algorithms and routines to address such questions. Possible applications lie in the study of mathematical conjectures where exact mathematical results are required (e.g., stability of dynamical systems); or in more applied questions, such as the automatic generation of efficient and reliable numerical software for function evaluation. On a complementary viewpoint, numerical safety is also critical in robust space mission design, where guidance and control algorithms become more complex in the context of increased satellite autonomy. We will pursue our collaboration with specialists of that area whose questions bring us interesting focus on relevant issues.

3.1.2. Floating-point computing

Floating-point arithmetic is currently undergoing a major evolution, in particular with the recent advent of a greater diversity of available precisions on a same system (from 8 to 128 bits) and of coarser-grained floating-point hardware instructions. This new arithmetic landscape raises important issues at the various levels of computing, that we will address along the following three directions.

3.1.2.1. Floating-point algorithms, properties, and standardization

One of our targets is the design of building blocks of computing (e.g., algorithms for the basic operations and functions, and algorithms for complex or double-word arithmetic). Establishing properties of these building blocks (e.g., the absence of "spurious" underflows/overflows) is also important. The IEEE 754 standard on floating-point arithmetic (whose next version, a rather minor revision, will be released soon) will have to undergo a major revision within a few years: first because advances in technology or new needs make some of its features obsolete, and because new features need standardization. We aim at playing a leading role in the preparation of the next standard.

3.1.2.2. Error bounds

We will pursue our studies in rounding error analysis, in particular for the "low precision–high dimension" regime, where traditional analyses become ineffective and where improved bounds are thus most needed. For this, the structure of both the data and the errors themselves will have to be exploited. We will also investigate the impact of mixed-precision and coarser-grained instructions (such as small matrix products) on accuracy analyses.

3.1.2.3. High performance kernels

Most directions in the team are concerned with optimized and high performance implementations. We will pursue our efforts concerning the implementation of well optimized floating-point kernels, with an emphasis on numerical quality, and taking into account the current evolution in computer architectures (the increasing width of SIMD registers, and the availability of low precision formats). We will focus on computing kernels used within other axes in the team such as, for example, extended precision linear algebra routines within the FPLLL and HPLLL libraries.

3.2. Lattices: algorithms and cryptology

We intend to strengthen our assessment of the cryptographic relevance of problems over lattices, and to broaden our studies in two main (complementary) directions: hardness foundations and advanced functionalities.

3.2.1. Hardness foundations

Recent advances in cryptography have broaden the scope of encryption functionalities (e.g., encryption schemes allowing to compute over encrypted data or to delegate partial decryption keys). While simple variants (e.g., identity-based encryption) are already practical, the more advanced ones still lack efficiency. Towards reaching practicality, we plan to investigate simpler constructions of the fundamental building blocks (e.g., pseudorandom functions) involved in these advanced protocols. We aim at simplifying known constructions based on standard hardness assumptions, but also at identifying new sources of hardness from which simple constructions that are naturally suited for the aforementioned advanced applications could be obtained (e.g., constructions that minimize critical complexity measures such as the depth of evaluation). Understanding the core source of hardness of today's standard hard algorithmic problems is an interesting direction as it could lead to new hardness assumptions (e.g., tweaked version of standard ones) from which we could derive much more efficient constructions. Furthermore, it could open the way to completely different constructions of advanced primitives based on new hardness assumptions.

3.2.2. Cryptanalysis

Lattice-based cryptography has come much closer to maturity in the recent past. In particular, NIST has started a standardization process for post-quantum cryptography, and lattice-based proposals are numerous and competitive. This dramatically increases the need for cryptanalysis: Do the underlying hard problems suffer from structural weaknesses? Are some of the problems used easy to solve, e.g., asymptotically? Are the chosen concrete parameters meaningful for concrete cryptanalysis? In particular, how secure would they be if all the known algorithms and implementations thereof were pushed to their limits? How would these concrete performances change in case (full-fledged) quantum computers get built?

On another front, the cryptographic functionalities reachable under lattice hardness assumptions seem to get closer to an intrinsic ceiling. For instance, to obtain cryptographic multilinear maps, functional encryption and indistinguishability obfuscation, new assumptions have been introduced. They often have a lattice flavour, but are far from standard. Assessing the validity of these assumptions will be one of our priorities in the mid-term.

3.2.3. Advanced cryptographic primitives

In the design of cryptographic schemes, we will pursue our investigations on functional encryption. Despite recent advances, efficient solutions are only available for restricted function families. Indeed, solutions for general functions are either way too inefficient for pratical use or they rely on uncertain security foundations like the existence of circuit obfuscators (or both). We will explore constructions based on well-studied hardness assumptions and which are closer to being usable in real-life applications. In the case of specific functionalities, we will aim at more efficient realizations satisfying stronger security notions.

Another direction we will explore is multi-party computation via a new approach exploiting the rich structure of class groups of quadratic fields. We already showed that such groups have a positive impact in this field by designing new efficient encryption switching protocols from the additively homomorphic encryption we introduced earlier. We want to go deeper in this direction that raises interesting questions such as how to design efficient zero-knowledge proofs for groups of unknown order, how to exploit their structure in the context of 2-party cryptography (such as two-party signing) or how to extend to the multi-party setting.

In the context of the PROMETHEUS H2020 project, we will keep seeking to develop new quantum-resistant privacy-preserving cryptographic primitives (group signatures, anonymous credentials, e-cash systems, etc). This includes the design of more efficient zero-knowledge proof systems that can interact with lattice-based cryptographic primitives.

3.3. Algebraic computing and high performance kernels

The connections between algorithms for structured matrices and for polynomial matrices will continue to be developed, since they have proved to bring progress to fundamental questions with applications throughout computer algebra. The new fast algorithm for the bivariate resultant opens an exciting area of research which should produce improvements to a variety of questions related to polynomial elimination. Obviously, we expect to produce results in that area.

For definite summation and integration, we now have fast algorithms for single integrals of general functions and sequences and for multiple integrals of rational functions. The long-term objective of that part of computer algebra is an efficient and general algorithm for multiple definite integration and summation of general functions and sequences. This is the direction we will take, starting with single definite sums of general functions and sequences (leading in particular to a faster variant of Zeilberger's algorithm). We also plan to investigate geometric issues related to the presence of apparent singularities and how they seem to play a role in the complexity of the current algorithms.

AROMATH Project-Team

3. Research Program

3.1. High order geometric modeling

The accurate description of shapes is a long standing problem in mathematics, with an important impact in many domains, inducing strong interactions between geometry and computation. Developing precise geometric modeling techniques is a critical issue in CAD-CAM. Constructing accurate models, that can be exploited in geometric applications, from digital data produced by cameras, laser scanners, observations or simulations is also a major issue in geometry processing. A main challenge is to construct models that can capture the geometry of complex shapes, using few parameters while being precise.

Our first objective is to develop methods, which are able to describe accurately and in an efficient way, objects or phenomena of geometric nature, using algebraic representations.

The approach followed in CAGD, to describe complex geometry is based on parametric representations called NURBS (Non Uniform Rational B-Spline). The models are constructed by trimming and gluing together high order patches of algebraic surfaces. These models are built from the so-called B-Spline functions that encode a piecewise algebraic function with a prescribed regularity at knots. Although these models have many advantages and have become the standard for designing nowadays CAD models, they also have important drawbacks. Among them, the difficulty to locally refine a NURBS surface and also the topological rigidity of NURBS patches that imposes to use many such patches with trims for designing complex models, with the consequence of the appearing of cracks at the seams. To overcome these difficulties, an active area of research is to look for new blending functions for the representation of CAD models. Some examples are the so-called T-Splines, LR-Spline blending functions, or hierarchical splines, that have been recently devised in order to perform efficiently local refinement. An important problem is to analyze spline spaces associated to general subdivisions, which is of particular interest in higher order Finite Element Methods. Another challenge in geometric modeling is the efficient representation and/or reconstruction of complex objects, and the description of computational domains in numerical simulation. To construct models that can represent efficiently the geometry of complex shapes, we are interested in developing modeling methods, based on alternative constructions such as skeleton-based representations. The change of representation, in particular between parametric and implicit representations, is of particular interest in geometric computations and in its applications in CAGD.

We also plan to investigate adaptive hierarchical techniques, which can locally improve the approximation of a shape or a function. They shall be exploited to transform digital data produced by cameras, laser scanners, observations or simulations into accurate and structured algebraic models.

The precise and efficient representation of shapes also leads to the problem of extracting and exploiting characteristic properties of shapes such as symmetry, which is very frequent in geometry. Reflecting the symmetry of the intended shape in the representation appears as a natural requirement for visual quality, but also as a possible source of sparsity of the representation. Recognizing, encoding and exploiting symmetry requires new paradigms of representation and further algebraic developments. Algebraic foundations for the exploitation of symmetry in the context of non linear differential and polynomial equations are addressed. The intent is to bring this expertise with symmetry to the geometric models and computations developed by AROMATH.

3.2. Robust algebraic-geometric computation

In many problems, digital data are approximated and cannot just be used as if they were exact. In the context of geometric modeling, polynomial equations appear naturally, as a way to describe constraints between the unknown variables of a problem. *An important challenge is to take into account the input error in order to*

develop robust methods for solving these algebraic constraints. Robustness means that a small perturbation of the input should produce a controlled variation of the output, that is forward stability, when the input-output map is regular. In non-regular cases, robustness also means that the output is an exact solution, or the most coherent solution, of a problem with input data in a given neighborhood, that is backward stability.

Our second long term objective is to develop methods to robustly and efficiently solve algebraic problems that occur in geometric modeling.

Robustness is a major issue in geometric modeling and algebraic computation. Classical methods in computer algebra, based on the paradigm of exact computation, cannot be applied directly in this context. They are not designed for stability against input perturbations. New investigations are needed to develop methods, which integrate this additional dimension of the problem. Several approaches are investigated to tackle these difficulties.

One relies on linearization of algebraic problems based on "elimination of variables" or projection into a space of smaller dimension. Resultant theory provides strong foundation for these methods, connecting the geometric properties of the solutions with explicit linear algebra on polynomial vector spaces, for families of polynomial systems (e.g., homogeneous, multi-homogeneous, sparse). Important progresses have been made in the last two decades to extend this theory to new families of problems with specific geometric properties. Additional advances have been achieved more recently to exploit the syzygies between the input equations. This approach provides matrix based representations, which are particularly powerful for approximate geometric computation on parametrized curves and surfaces. They are tuned to certain classes of problems and an important issue is to detect and analyze degeneracies and to adapt them to these cases.

A more adaptive approach involves linear algebra computation in a hierarchy of polynomial vector spaces. It produces a description of quotient algebra structures, from which the solutions of polynomial systems can be recovered. This family of methods includes Gröbner Basis, which provides general tools for solving polynomial equations. Border Basis is an alternative approach, offering numerically stable methods for solving polynomial equations with approximate coefficients. An important issue is to understand and control the numerical behavior of these methods as well as their complexity and to exploit the structure of the input system.

In order to compute "only" the (real) solutions of a polynomial system in a given domain, duality techniques can also be employed. They consist in analyzing and adding constraints on the space of linear forms which vanish on the polynomial equations. Combined with semi-definite programming techniques, they provide efficient methods to compute the real solutions of algebraic equations or to solve polynomial optimization problems. The main issues are the completness of the approach, their scalability with the degree and dimension and the certification of bounds.

Singular solutions of polynomial systems can be analyzed by computing differentials, which vanish at these points. This leads to efficient deflation techniques, which transform a singular solution of a given problem into a regular solution of the transformed problem. These local methods need to be combined with more global root localisation methods.

Subdivision methods are another type of methods which are interesting for robust geometric computation. They are based on exclusion tests which certify that no solution exists in a domain and inclusion tests, which certify the uniqueness of a solution in a domain. They have shown their strength in addressing many algebraic problems, such as isolating real roots of polynomial equations or computing the topology of algebraic curves and surfaces. The main issues in these approaches is to deal with singularities and degenerate solutions.

CAIRN Project-Team

3. Research Program

3.1. Panorama

The development of complex applications is traditionally split in three stages: a theoretical study of the algorithms, an analysis of the target architecture and the implementation. When facing new emerging applications such as high-performance, low-power and low-cost mobile communication systems or smart sensor-based systems, it is mandatory to strengthen the design flow by a joint study of both algorithmic and architectural issues.

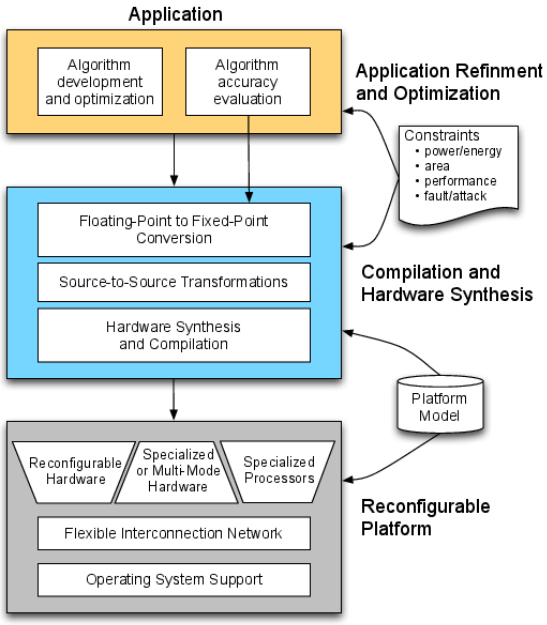


Figure 1. CAIRN*'s general design flow and related research themes*

Figure 1 shows the global design flow we propose to develop. This flow is organized in levels corresponding to our three research themes: application optimization (new algorithms, fixed-point arithmetic, advanced representations of numbers), architecture optimization (reconfigurable and specialized hardware, application-specific processors, arithmetic operators and functions), and stepwise refinement and code generation (code transformations, hardware synthesis, compilation).
In the rest of this part, we briefly describe the challenges concerning **new reconfigurable platforms** in Section 3.2 and the issues on **compiler and synthesis tools** related to these platforms in Section 3.3 .

3.2. Reconfigurable Architecture Design

Nowadays, FPGAs are not only suited for application specific algorithms, but also considered as fully-featured computing platforms, thanks to their ability to accelerate massively parallelizable algorithms much faster than their processor counterparts [69]. They can also be reconfigured dynamically. At runtime, partially reconfigurable regions of the logic fabric can be reconfigured to implement a different task, which allows for a better resource usage and adaptation to the environment. Dynamically reconfigurable hardware can also cope with hardware errors by relocating some of its functionalities to another, sane, part of the logic fabric. It could also provide support for a multi-tasked computation flow where hardware tasks are loaded on-demand at runtime. Nevertheless, current design flows of FPGA vendors are still limited by the use of one partial bitstream for each reconfigurable region and for each design. These regions are defined at design time and it is not possible to use only one bitstream for multiple reconfigurable regions nor multiple chips. The multiplicity of such bitstreams leads to a significant increase in memory. Recent research has been conducted in the domain of task relocation on a reconfigurable fabric. All related work has been conducted on architectures from commercial vendors (e.g., Xilinx, Altera) which share the same limitations: the inner details of the bitstream are not publicly known, which limits applicability of the techniques. To circumvent this issue, most dynamic reconfiguration techniques are either generating multiple bitstreams for each location [53] or implementing an online filter to relocate the tasks [63]. Both of these techniques still suffer from memory footprint and from the online complexity of task relocation.

Increasing the level and grain of reconfiguration is a solution to counterbalance the FPGA penalties. Coarse-grained reconfigurable architectures (CGRA) provide operator-level configurable functional blocks and word-level datapaths [70], [58], [68]. Compared to FPGA, they benefit from a massive reduction in configuration memory and configuration delay, as well as for routing and placement complexity. This in turns results in an improvement in the computation volume over energy cost ratio, although with a loss of flexibility compared to bit-level operations. Such constraints have been taken into account in the design of DART[9], Adres [66] or polymorphous computing fabrics[11]. These works have led to commercial products such as the PACT/XPP [52] or Montium from Recore systems, without however a real commercial success yet. Emerging platforms like Xilinx/Zynq or Intel/Altera are about to change the game.

In the context of emerging heterogenous multicore architecture, CAIRN advocates for associating general-purpose processors (GPP), flexible network-on-chip and coarse-grain or fine-grain dynamically reconfigurable accelerators. We leverage our skills on microarchitecture, reconfigurable computing, arithmetic, and low-power design, to discover and design such architectures with a focus on: reduced energy per operation; improved application performance through acceleration; hardware flexibility and self-adaptive behavior; tolerance to faults, computing errors, and process variation; protections against side channel attacks; limited silicon area overhead.

3.3. Compilation and Synthesis for Reconfigurable Platforms

In spite of their advantages, reconfigurable architectures, and more generally hardware accelerators, lack efficient and standardized compilation and design tools. As of today, this still makes the technology impractical for large-scale industrial use. Generating and optimizing the mapping from high-level specifications to reconfigurable hardware platforms are therefore key research issues, which have received considerable interest over the last years [56], [71], [67], [65], [64]. In the meantime, the complexity (and heterogeneity) of these platforms has also been increasing quite significantly, with complex heterogeneous multi-cores architectures becoming a *de facto* standard. As a consequence, the focus of designers is now geared toward optimizing overall system-level performance and efficiency [62]. Here again, existing tools are not well suited, as they fail at providing a unified programming view of the programmable and/or reconfigurable components implemented on the platform.

In this context, we have been pursuing our efforts to propose tools whose design principles are based on a tight coupling between the compiler and the target hardware architectures. We build on the expertise of the team members in High Level Synthesis (HLS) [5], ASIP optimizing compilers [12] and automatic parallelization for massively parallel specialized circuits [2]. We first study how to increase the efficiency of standard programmable processors by extending their instruction set to speed-up compute intensive kernels. Our focus is on efficient and exact algorithms for the identification, selection and scheduling of such instructions [6]. We address compilation challenges by borrowing techniques from high-level synthesis, optimizing compilers and automatic parallelization, especially when dealing with nested loop kernels. In addition, and independently of the scientific challenges mentioned above, proposing such flows also poses significant software engineering issues. As a consequence, we also study how leading edge software engineering techniques (Model Driven Engineering) can help the Computer Aided Design (CAD) and optimizing compiler communities prototyping new research ideas [4].

Efficient implementation of multimedia and signal processing applications (in software for DSP cores or as special-purpose hardware) often requires, for reasons related to cost, power consumption or silicon area constraints, the use of fixed-point arithmetic, whereas the algorithms are usually specified in floating-point arithmetic. Unfortunately, fixed-point conversion is very challenging and time-consuming, typically demanding up to 50% of the total design or implementation time. Thus, tools are required to automate this conversion. For hardware or software implementation, the aim is to optimize the fixed-point specification. The implementation cost is minimized under a numerical accuracy or an application performance constraint. For DSP-software implementation, methodologies have been proposed [7] to achieve fixed-point conversion. For hardware implementation, the best results are obtained when the word-length optimization process is coupled with the high-level synthesis [59]. Evaluating the effects of finite precision is one of the major and often the most time consuming step while performing fixed-point refinement. Indeed, in the word-length optimization process, the numerical accuracy is evaluated as soon as a new word-length is tested, thus, several times per iteration of the optimization process. Classical approaches are based on fixed-point simulations [60]. Leading to long evaluation times, they can hardly be used to explore the design space. Therefore, our aim is to propose closed-form expressions of errors due to fixed-point approximations that are used by a fast analytical framework for accuracy evaluation [10].

CAMBIUM Project-Team

3. Research Program

3.1. Research Directions

Our research proposal is organized along three main axes, namely **programming language design and implementation**, **concurrency**, and **program verification**. These three areas have strong connections. For instance, the definition and implementation of Multicore OCaml intersects the first two axes, whereas creating verification technology for Multicore OCaml programs intersects the last two.

In short, the "programming language design and implementation" axis includes:

- The search for richer type disciplines, in an effort to make our programming languages safer and more expressive. Two domains, namely modules and effects, appear of particular interest. In addition, we view type inference as an important cross-cutting concern.
- The continued evolution of OCaml. The major evolutions that we envision in the medium term are the integration of Multicore OCaml, the addition of modular implicits, and a redesign of the type-checker.
- Research on refactoring and program transformations.

The "concurrency" axis includes:

- Research on weak memory models, including axiomatic models, operational models, and event-structure models.
- Research on the Multicore OCaml memory model. This might include proving that the axiomatic and operational presentations of the model agree; testing the Multicore OCaml implementation to ensure that it conforms to the model; and extending the model with new features, should the need arise.

The "program verification" axis includes:

- The continued evolution of CompCert.
- Building new verified tools, such as verified compilers for domain-specific languages, verified components for the Coq type-checker, and so on.
- Verifying algorithms and data structures implemented in OCaml and in Multicore OCaml and enriching Separation Logic with new features, if needed, to better support this activity.
- The continued development of tools for TLA+.

CAMUS Project-Team

3. Research Program

3.1. Research Directions

The various objectives we are expecting to reach are directly related to the search of adequacy between the sofware and the new multicore processors evolution. They also correspond to the main research directions suggested by Hall, Padua and Pingali in [56]. Performance, correctness and productivity must be the users' perceived effects. They will be the consequences of research works dealing with the following issues:

- Issue 1: Static Parallelization and Optimization
- Issue 2: Profiling and Execution Behavior Modeling
- Issue 3: Dynamic Program Parallelization and Optimization, Virtual Machine
- Issue 4: Proof of Program Transformations for Multicores

The development of efficient and correct applications for multicore processors requires stepping in every application development phase, from the initial conception to the final run.

Upstream, all potential parallelism of the application has to be exhibited. Here static analysis and transformation approaches (issue 1) must be performed, resulting in *multi-parallel* intermediate code advising the running virtual machine about all the parallelism that can be taken advantage of. However the compiler does not have much knowledge about the execution environment. It obviously knows the instruction set, it can be aware of the number of available cores, but it does not know the actual available resources at any time during the execution (memory, number of free cores, etc.).

That is the reason why a "virtual machine" mechanism will have to adapt the application to the resources (issue 3). Moreover the compiler will be able to take advantage only of a part of the parallelism induced by the application. Indeed some program information (variables values, accessed memory adresses, etc.) being available only at runtime, another part of the available parallelism will have to be generated on-the-fly during the execution, here also, thanks to a dynamic mechanism.

This on-the-fly parallelism extraction will be performed using speculative behavior models (issue 2), such models allowing to generate speculative parallel code (issue 3). Between our behavior modeling objectives, we can add the behavior monitoring, or profiling, of a program version. Indeed, the complexity of current and future architectures avoids assuming an optimal behavior regarding a given program version. A monitoring process will make it possible to select on-the-fly the best parallelization.

These different parallelization steps are schematized in figure 1 .

Our project relies on the conception of a production chain for efficient execution of an application on a multicore architecture. Each link of this chain has to be formally verified in order to ensure correctness as well as efficiency. More precisely, it has to be ensured that the compiler produces a correct intermediate code, and that the virtual machine actually performs the parallel execution semantically equivalent to the source code: every transformation applied to the application, either statically by the compiler or dynamically by the virtual machine, must preserve the initial semantics. This must be proved formally (issue 4).

In the following, those different issues are detailed while forming our global, long term vision of what has to be done.

3.2. Static Parallelization and Optimization

Participants: Vincent Loechner, Philippe Clauss, Éric Violard, Cédric Bastoul, Arthur Charguéraud, Bérenger Bramas, Harenome Ranaivoarivony-Razanajato.

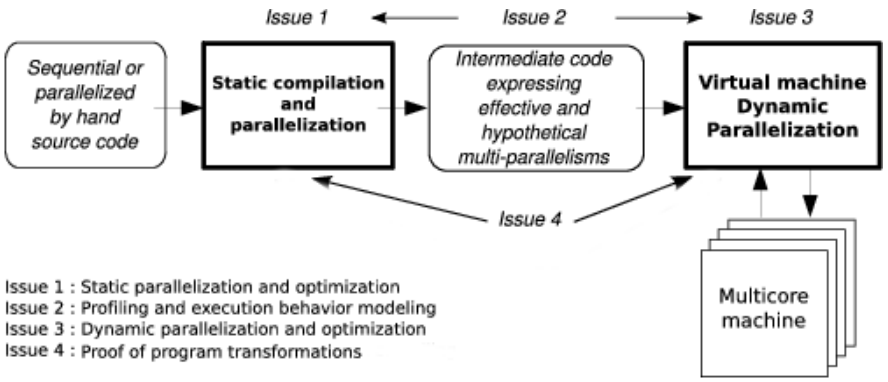


Figure 1. Steps for Automatic parallelization on multicore architectures.

Static optimizations, from source code at compile time, benefit from two decades of research in automatic parallelization: many works address the parallelization of loop nests accessing multi-dimensional arrays, and these works are now mature enough to generate efficient parallel code [53]. Low-level optimizations, in the assembly code generated by the compiler, have also been extensively dealt with for single-core and require few adaptations to support multicore architectures. Concerning multicore specific parallelization, we propose to explore two research directions to take full advantage of these architectures: adapting parallelization to multicore architectures and expressing many potential parallelisms.

3.3. Profiling and Execution Behavior Modeling

Participants: Alain Ketterlin, Philippe Clauss, Salwa Kobeissi.

The increasing complexity of programs and hardware architectures makes it ever harder to characterize beforehand a given program's run time behavior. The sophistication of current compilers and the variety of transformations they are able to apply cannot hide their intrinsic limitations. As new abstractions like transactional memories appear, the dynamic behavior of a program strongly conditions its observed performance. All these reasons explain why empirical studies of sequential and parallel program executions have been considered increasingly relevant. Such studies aim at characterizing various facets of one or several program runs, *e.g.*, memory behavior, execution phases, etc. In some cases, such studies characterize more the compiler than the program itself. These works are of tremendous importance to highlight all aspects that escape static analysis, even though their results may have a narrow scope, due to the possible incompleteness of their input data sets.

3.4. Dynamic Parallelization and Optimization, Virtual Machine

Participants: Philippe Clauss, Salwa Kobeissi, Jens Gustedt, Alain Ketterlin, Muthena Abdul-Wahab, Daniel Salas, Bérenger Bramas.

Dynamic parallelization and optimization has become essential with the advent of the new multicore architectures. When using a dynamic scheme, the performed instructions are not only dedicated to the application functionalities, but also to its control and its transformation, and so in its own interest. Behaving like a computer virus, such a scheme should rather be qualified as a "vitamin". It perfectly knows the current characteristics of the execution environment and owns some qualitative information thanks to a behavior modeling process (issue 2). It provides a significant optimization ability compared to a static compiler, while observing the evolution of the availability of live resources.

3.5. Proof of Program Transformations for Multicores

Participants: Éric Violard, Alain Ketterlin, Julien Narboux, Nicolas Magaud, Arthur Charguéraud.

Our main objective consists in certifying the critical modules of our optimization tools (the compiler and the virtual machine). First we will prove the main loop transformation algorithms which constitute the core of our system.

The optimization process can be separated into two stages: the transformations consisting in optimizing the sequential code and in exhibiting parallelism, and those consisting in optimizing the parallel code itself. The first category of optimizations can be proved within a sequential semantics. For the other optimizations, we need to work within a concurrent semantics. We expect the first stage of optimization to produce data-race free code. For the second stage of optimization we will first assume that the input code is data-race free. We will prove those transformations using Appel's concurrent separation logic [57]. Proving transformations involving programs which are not data-race free will constitute a longer term research goal.

CARAMBA Project-Team

3. Research Program

3.1. The Extended Family of the Number Field Sieve

The Number Field Sieve (NFS) has been the leading algorithm for factoring integers for more than 20 years, and its variants have been used to set records for discrete logarithms in finite fields. It is reasonable to understand NFS as a framework that can be used to solve various sorts of problems. Factoring integers and computing discrete logarithms are the most prominent for the cryptographic observer, but the same framework can also be applied to the computation of class groups.

The state of the art with NFS is built from numerous improvements of its inner steps. In terms of algorithmic improvements, the recent research activity on the NFS family has been rather intense. Several new algorithms have been discovered since 2014, notably for non-prime fields, and their practical reach has been demonstrated by actual experiments.

The algorithmic contributions of the CARAMBA members to NFS would hardly be possible without access to a dependable software implementation. To this end, members of the CARAMBA team have been developing the Cado-NFS software suite since 2007. Cado-NFS is now the most widely visible open-source implementation of NFS, and is a crucial platform for developing prototype implementations for new ideas for the many sub-algorithms of NFS. Cado-NFS is free software (LGPL) and follows an open development model, with publicly accessible development repository and regular software releases. Competing free software implementations exist, such as `msieve`, developed by J. Papadopoulos (whose last commit is from August 2018). In Lausanne, T. Kleinjung develops his own code base, which is unfortunately not public.

The work plan of CARAMBA on the topic of the Number Field Sieve algorithm and its cousins includes the following aspects:

- Pursue the work on NFS, which entails in particular making it ready to tackle larger challenges. Several of the important computational steps of NFS that are currently identified as stumbling blocks will require algorithmic advances and implementation improvements. We will illustrate the importance of this work by computational records.
- Work on the specific aspects of the computation of discrete logarithms in finite fields.
- As a side topic, the application of the broad methodology of NFS to the treatment of "ideal lattices" and their use in cryptographic proposals based on Euclidean lattices is also relevant.

3.2. Algebraic Curves for Cryptology

The challenges associated with algebraic curves in cryptology are diverse, because of the variety of mathematical objects to be considered. These challenges are also connected to each other. On the cryptographic side, efficiency matters. With the standardization of TLS 1.3 in 2018 [34], the curves x25519 and x448 have entered the base specification of standard. These curves were designed by academia and offer an excellent compromise between efficiency and security.

On the cryptanalytic side, the discrete logarithm problem on (Jacobians of) curves has resisted all attempts for many years. Among the currently active topics, the decomposition algorithms raise interesting problems related to polynomial system solving, as do attempts to solve the discrete logarithm problem on curves defined over binary fields. In particular, while it is generally accepted that the so-called Koblitz curves (base field extensions of curves defined over $\mathrm{GF}(2)$) are likely to be a weak class among the various curve choices, no concrete attack supports this claim fully.

The research objectives of CARAMBA on the topic of algebraic curves for cryptology are as follows:

- Work on the practical realization of some of the rich mathematical theory behind algebraic curves. In particular, some of the fundamental mathematical objects have potentially important connections to the broad topic of cryptology: Abel-Jacobi map, Theta functions, computation of isogenies, computation of endomorphisms, complex multiplication.
- Improve the point counting algorithms so as to be able to tackle larger problems. This includes significant work connected to polynomial systems.
- Seek improvements on the computation of discrete logarithms on curves, including by identifying weak instances of this problem.

3.3. Symmetric Cryptography

Since the recruiting of Marine Minier in September 2016 as a Professor at the Université de Lorraine, and of Virginie Lallemand as a CNRS researcher in October 2018, a new research domain has emerged in the CARAMBA team: symmetric key cryptology. Accompanied in this adventure by non-permanent team members, we are tackling problems related to both design and analysis. A large part of our recent researches has been motivated by the Lightweight Cryptography Standardization Process of the NIST [0] that embodies a crucial challenge of the last decade: finding ciphers that are suitable for resource-constrained devices.

On a general note, the working program of CARAMBA in symmetric cryptography is defined as follows:

- Develop automatic tools based on constraint programming to help finding optimum attack parameters. The effort will be focused on the AES standard and on recent lightweight cipher proposals.
- Contribute to the security and performance analysis effort required to sort out the candidates for the NIST Lightweight Cryptography Standardization Process.
- Study how to protect services execution on dedicated platforms using white-box cryptography and software obfuscation methods.

3.4. Computer Arithmetic

Computer arithmetic is part of the common background of all team members, and is naturally ubiquitous in our application domains. However involved the mathematical objects considered may be, dealing with them first requires to master more basic objects: integers, finite fields, polynomials, and real and complex floating-point numbers. Libraries such as GNU MP, GNU MPFR, GNU MPC do an excellent job for these, both for small and large sizes (we rarely, if ever, focus on small-precision floating-point data, which explains our lack of mention of libraries relevant to it).

Most of our involvement in subjects related to computer arithmetic is to be understood in connection to our applications to the Number Field Sieve and to abelian varieties. As such, much of the research work we envision will appear as side-effects of developments in these contexts. On the topic of arithmetic work *per se*:

- We will seek algorithmic and practical improvements to the most basic algorithms. That includes for example the study of advanced algorithms for integer multiplication, and their practical reach.
- We will continue to work on the arithmetic libraries in which we have crucial involvement, such as GNU MPFR, GNU MPC, GF2X, MPFQ, and also GMP-ECM.

3.5. Polynomial Systems

Systems of polynomial equations have been part of the cryptographic landscape for quite some time, with applications to the cryptanalysis of block and stream ciphers, as well as multivariate cryptographic primitives.

[0] National Institute of Standard and Technology.

Polynomial systems arising from cryptology are usually not generic, in the sense that they have some distinct structural properties, such as symmetries, or bi-linearity for example. During the last decades, several results have shown that identifying and exploiting these structures can lead to dedicated Gröbner basis algorithms that can achieve large speedups compared to generic implementations [29], [28].

Solving polynomial systems is well done by existing software, and duplicating this effort is not relevant. However we develop test-bed open-source software for ideas relevant to the specific polynomial systems that arise in the context of our applications. The TinyGB software is our platform to test new ideas.

We aim to work on the topic of polynomial system solving in connection with our involvement in the aforementioned topics.

- We have high expertise on Elliptic Curve Cryptography in general. On the narrower topic of the Elliptic Curve Discrete Logarithm Problem on small characteristic finite fields, the highly structured polynomial systems that are involved match well our expertise on the topic of polynomial systems. Once a very hot topic in 2015, activity on this precise problem seems to have slowed down. Yet, the conjunction of skills that we have may lead to results in this direction in the future.
- The hiring of Marine Minier is likely to lead the team to study particular polynomial systems in contexts related to symmetric key cryptography.
- More centered on polynomial systems *per se*, we will mainly pursue the study of the specificities of the polynomial systems that are strongly linked to our targeted applications, and for which we have significant expertise [29], [28]. We also want to see these recent results provide practical benefits compared to existing software, in particular for systems relevant for cryptanalysis.

CASCADE Project-Team

3. Research Program

3.1. Quantum-Safe Cryptography

The security of almost all public-key cryptographic protocols in use today relies on the presumed hardness of problems from number theory such as factoring and computing discrete logarithms. This is problematic because these problems have very similar underlying structure, and its unforeseen exploit can render all currently used public-key cryptography insecure. This structure was in fact exploited by Shor to construct efficient quantum algorithms that break all hardness assumptions from number theory that are currently in use. And so naturally, an important area of research is to build provably secure protocols based on mathematical problems that are unrelated to factoring and discrete log. One of the most promising directions in this line of research is using lattice problems as a source of computational hardness, which also offer features that other alternative public-key cryptosystems (such as MQ-based, code-based or hash-based schemes) cannot provide.

3.2. Advanced Encryption

Fully Homomorphic Encryption (FHE) has become a very active research area since 2009, when IBM announced the discovery of a FHE scheme by Craig Gentry. FHE allows to perform any computation on encrypted data, yielding the result encrypted under the same key. This enables outsourcing computation in the Cloud, on encrypted data, so the Cloud provider does not learn any information. However, FHE does not allow to share the result.

Functional encryption is another recent tool that allows an authority to deliver functional decryption keys, for any function f of his choice, so that when applied to the encryption of a message m, the functional decryption key yields $f(m)$. Since m can be a large vector, f can be an aggregation or statistical function: on encrypted data, one can get the result $f(m)$ in clear.

While this functionality has initially been defined in theory, our team has been very active in designing concrete instantiations for practical purposes.

3.3. Security amidst Concurrency on the Internet

Cryptographic protocols that are secure when executed in isolation can become completely insecure when multiple such instances are executed concurrently (as is unavoidable on the Internet) or when used as a part of a larger protocol. For instance, a man-in-the-middle attacker participating in two simultaneous executions of a cryptographic protocol might use messages from one of the executions in order to compromise the security of the second – Lowe's attack on the Needham-Schroeder authentication protocol and Bleichenbacher's attack on SSL work this way. Our research addresses security amidst concurrent executions in secure computation and key exchange protocols.

Secure computation allows several mutually distrustful parties to collaboratively compute a public function of their inputs, while providing the same security guarantees as if a trusted party had performed the computation. Potential applications for secure computation include anonymous voting, privacy-preserving auctions and data-mining. Our recent contributions on this topic include

1. new protocols for secure computation in a model where each party interacts only once, with a single centralized server; this model captures communication patterns that arise in many practical settings, such as that of Internet users on a website, and
2. efficient constructions of universally composable commitments and oblivious transfer protocols, which are the main building blocks for general secure computation.

In key exchange protocols, we are actively involved in designing new password-authenticated key exchange protocols, as well as the analysis of the widely-used SSL/TLS protocols.

3.4. Electronic Currencies and the Blockchain

Electronic cash (e-cash) was first proposed in the 1980s but has never been deployed on a large scale. Other means of digital payments are instead largely replacing physical cash, but they do not respect the citizens' right to privacy, which includes their right of anonymous payments of moderate sums. Recently, so-called decentralized currencies, such as Bitcoin, have become a third type of payments in addition to physical cash, and card and other (non-anonymous) electronic payments. The continuous growth of popularity and usage of this new kind of currencies, also called "cryptocurrencies", have triggered a renewed interest in cryptographic e-cash.

On the one hand, our group investigates "centralized" e-cash, in keeping with the current economic model that has money be issued by (central) banks (while cryptocurrencies use money distribution as an incentive for participation in the system, on which its stability hinges). Of particular interest among centralized e-cash schemes is transferable e-cash, which allows users to transfer coins between each other without interacting with a third party (or the blockchain). Existing efficient e-cash schemes are not transferable, as they require coins to be deposited at the bank after having been used in a payment. Our goal is to propose efficient transferable e-cash schemes.

Another direction concerns (decentralized) cryptocurrencies, whose adoption has grown tremendously over the last few years. While in Bitcoin all transactions are publicly posted on the so-called "blockchain", other cryptocurrencies such as *Zcash* respect user privacy, whose security guarantees we have analyzed. Apart from privacy, two pressing challenges for cryptocurrencies, and blockchains in general, are sustainability and scalability. Regarding the former, we are addressing the electricity waste caused by the concept of "proof of work" used by all major cryptocurrencies by proposing alternatives; for the latter, we are working on proposals that avoid the need for all data having to be stored on the blockchain forever.

Blockchains have meanwhile found many other applications apart from electronic money. Together with Microsoft Research, our group investigates decentralized means of authentication that uses cryptography to guarantee privacy.

CASH Project-Team

3. Research Program

3.1. Definition of dataflow representations of parallel programs

In the last decades, several frameworks have emerged to design efficient compiler algorithms. The efficiency of all the optimizations performed in compilers strongly relies on effective *static analyses* and *intermediate representations*. Dataflow models are a natural intermediate representation for hardware compilers (HLS) and more generally for parallelizing compilers. Indeed, dataflow models capture task-level parallelism and can be mapped naturally to parallel architectures. In a way, a dataflow model is a partition of the computation into processes and a partition of the flow dependences into channels. This partitioning prepares resource allocation (which processor/hardware to use) and medium-grain communications.

The main goal of the CASH team is to provide efficient analyses and the optimizing compilation frameworks for dataflow programming models. The results of the team relies on programming languages and representation of programs in which parallelism and dataflow play a crucial role. This first research direction aims at defining these dataflow languages and intermediate representations, both from a practical perspective (syntax or structure), and from a theoretical point of view (semantics). This first research direction thus defines the models on which the other directions will rely. It is important to note that we do not restrict ourselves to a strict definition of dataflow languages: more generally, we are interested in the parallel languages in which dataflow synchronization plays a significant role.

Intermediate dataflow model. The intermediate dataflow model is a representation of the program that is adapted for optimization and scheduling. It will be obtained from the analysis of a (parallel or sequential) program and should at some point be used for compilation. The dataflow model must specify precisely its semantics and parallelism granularity. It must also be analyzable with polyhedral techniques, where powerful concepts exist to design compiler analysis, e.g., scheduling or resource allocation. Polyhedral Process Networks [58] extended with a module system could be a good starting point. But then, how to fit non-polyhedral parts of the program? A solution is to hide non-polyhedral parts into processes with a proper polyhedral abstraction. This organization between polyhedral and non-polyhedral processes will be a key aspect of our medium-grain dataflow model. The design of our intermediate dataflow model and the precise definition of its semantics will constitute a reliable basis to formally define and ensure the correctness of algorithms proposed by CASH: compilation, optimizations and analyses.

Dataflow programming languages. Dataflow paradigm has also been explored quite intensively in programming languages. Indeed, there exists a large panel of dataflow languages, whose characteristics differ notably, the major point of variability being the scheduling of agents and their communications. There is indeed a continuum from the synchronous dataflow languages like Lustre [42] or Streamit [55], where the scheduling is fully static, and general communicating networks like KPNs [45] or RVC-Cal [25] where a dedicated runtime is responsible for scheduling tasks dynamically, when they *can* be executed. These languages share some similarities with actor languages that go even further in the decoupling of processes by considering them as independent reactive entities. Another objective of the CASH team is to study dataflow programming *languages*, their semantics, their expressiveness, and their compilation. The specificity of the CASH team is that these languages will be designed taking into consideration the compilation using polyhedral techniques. In particular, we will explore which dataflow constructs are better adapted for our static analysis, compilation, and scheduling techniques. In practice we want to propose high-level primitives to express data dependency, this way the programmer can express parallelism in a dataflow way instead of the classical communication-oriented dependencies. The higher-level more declarative point of view makes programming easier but also give more optimization opportunities. These primitives will be inspired by the existing works in the polyhedral model framework, as well as dataflow languages, but also in the actors and active object languages [32] that nowadays introduce more and more dataflow primitives to enable data-driven interactions between agents, particularly with *futures* [30], [37].

3.1.1. Expected Impact

Consequently, the impact of this research direction is both the usability of our representation for static analyses and optimizations performed in Sections 3.2 and 3.3 , and the usability of its semantics to prove the correctness of these analyses.

3.1.2. Scientific Program

3.1.2.1. Short-term and ongoing activities.

We obtained preliminary experimental [22], [23], [38] and theoretical [43] results, exploring several aspects of dataflow models. The next step is to define accurately the intermediate dataflow model and to study existing programming and execution models:

- Define our medium-grain dataflow model. So far, a modular Polyhedral Process Networks appears as a natural candidate but it may need to be extended to be adapted to a wider range of applications. Precise semantics will have to be defined for this model to ensure the articulation with the activities discussed in Section 3.3 .
- Study precisely existing dataflow languages, their semantics, their programmability, and their limitations.

3.1.2.2. Medium-term activities.

In a second step, we will extend the existing results to widen the expressiveness of our intermediate representation and design new parallelism constructs. We will also work on the semantics of dataflow languages:

- Propose new stream programming models and a clean semantics where all kinds of parallelisms are expressed explicitly, and where all activities from code design to compilation and scheduling can be clearly expressed.
- Identify a core language that is rich enough to be representative of the dataflow languages we are interested in, but abstract and small enough to enable formal reasoning and proofs of correctness for our analyses and optimizations.

3.1.2.3. Long-term activities.

In a longer-term vision, the work on semantics, while remaining driven by the applications, would lead to to more mature results, for instance:

- Design more expressive dataflow languages and intermediate representations which would at the same time be expressive enough to capture all the features we want for aggressive HPC optimizations, and sufficiently restrictive to be (at least partially) statically analyzable at a reasonable cost.
- Define a module system for our medium-grain dataflow language. A program will then be divided into modules that can follow different compilation schemes and execution models but still communicate together. This will allow us to encapsulate a program that does not fit the polyhedral model into a polyhedral one and vice versa. Also, this will allow a compositional analysis and compilation, as opposed to global analysis which is limited in scalability.

3.2. Expressivity and Scalability of Static Analyses

The design and implementation of efficient compilers becomes more difficult each day, as they need to bridge the gap between *complex languages* and *complex architectures*. Application developers use languages that bring them close to the problem that they need to solve which explains the importance of high-level programming languages. However, high-level programming languages tend to become more distant from the hardware which they are meant to command.

In this research direction, we propose to design expressive and scalable static analyses for compilers. This topic is closely linked to Sections 3.1 and 3.3 since the design of an efficient intermediate representation is made while regarding the analyses it enables. The intermediate representation should be expressive enough to embed maximal information; however if the representation is too complex the design of scalable analyses will be harder.

The analyses we plan to design in this activity will of course be mainly driven by the HPC dataflow optimizations we mentioned in the preceding sections; however we will also target other kinds of analyses applicable to more general purpose programs. We will thus consider two main directions:

- Extend the applicability of the polyhedral model, in order to deal with HPC applications that do not fit totally in this category. More specifically, we plan to work on more complex control and also on complex data structures, like sparse matrices, which are heavily used in HPC.
- Design of specialized static analyses for memory diagnostic and optimization inside general purpose compilers.

For both activities, we plan to cross fertilize ideas coming from the abstract interpretation community as well as language design, dataflow semantics, and WCET estimation techniques.

Correct by construction analyses. The design of well-defined semantics for the chosen programming language and intermediate representation will allow us to show the correctness of our analyses. The precise study of the semantics of Section 3.1 will allow us to adapt the analysis to the characteristics of the language, and prove that such an adaptation is well founded. This approach will be applicable both on the source language and on the intermediate representation.

Such wellfoundedness criteria relatively to the language semantics will first be used to design our analyses, and then to study which extensions of the languages can be envisioned and analyzed safely, and which extensions (if any) are difficult to analyze and should be avoided. Here the correct identification of a core language for our formal studies (see Section 3.1) will play a crucial role as the core language should feature all the characteristics that might make the analysis difficult or incorrect.

Scalable abstract domains. We already have experience in designing low-cost semi relational abstract domains for pointers [50], [46], as well as tailoring static analyses for specialized applications in compilation [36], [54], Synchronous Dataflow scheduling [53], and extending the polyhedral model to irregular applications [21]. We also have experience in the design of various static verification techniques adapted to different programming paradigms.

3.2.1. Expected impact

The impact of this work is the significantly widened applicability of various tools/compilers related to parallelization: allow optimizations for a larger class of programs, and allow low-cost analysis that scale to very large programs.

We target both analysis for optimization and analysis to detect, or prove the absence of bugs.

3.2.2. Scientific Program

3.2.2.1. Short-term and ongoing activities.

Together with Paul Iannetta and Lionel Morel (INSA/CEA LETI), we are currently working on the *semantic rephrasing* of the polyhedral model [39]. The objective is to clearly redefine the key notions of the polyhedral model on general flowchart programs operating on arrays, lists and trees. We reformulate the algorithms that are performed to compute dependencies in a more semantic fashion, i.e. considering the program semantics instead of syntactical criteria. The next step is to express classical scheduling and code generation activities in this framework, in order to overcome the classical syntactic restrictions of the polyhedral model.

3.2.2.2. Medium-term activities.

In medium term, we want to extend the polyhedral model for more general data-structures like lists and sparse matrices. For that purpose, we need to find polyhedral (or other shapes) abstractions for non-array data-structures; the main difficulty is to deal with non-linearity and/or partial information (namely, over-approximations of the data layout, or over-approximation of the program behavior). This activity will rely on a formalization of the optimization activities (dependency computation, scheduling, compilation) in a more general Abstract-Interpretation based framework in order to make the approximations explicit.

At the same time, we plan to continue to work on scaling static analyses for general purpose programs, in the spirit of Maroua Maalej's PhD [47], whose contribution is a sequence of memory analyses inside production compilers. We already began a collaboration with Sylvain Collange (PACAP team of IRISA Laboratory) on the design of static analyses to optimize copies from the global memory of a GPU to the block kernels (to increase locality). In particular, we have the objective to design specialized analyses but with an explicit notion of cost/precision compromise, in the spirit of the paper [41] that tries to formalize the cost/precision compromise of interprocedural analyses with respect to a "context sensitivity parameter".

3.2.2.3. Long-term activities.

In a longer-term vision, the work on scalable static analyses, whether or not directed from the dataflow activities, will be pursued in the direction of large general-purpose programs.

An ambitious challenge is to find a generic way of adapting existing (relational) abstract domains within the Single Static Information [26] framework so as to improve their scalability. With this framework, we would be able to design static analyses, in the spirit of the seminal paper [31] which gave a theoretical scheme for classical abstract interpretation analyses.

We also plan to work on the interface between the analyses and their optimization clients inside production compilers.

3.3. Compiling and Scheduling Dataflow Programs

In this part, we propose to design the compiler analyses and optimizations for the *medium-grain* dataflow model defined in section 3.1 . We also propose to exploit these techniques to improve the compilation of dataflow languages based on actors. Hence our activity is split into the following parts:

- Translating a sequential program into a medium-grain dataflow model. The programmer cannot be expected to rewrite the legacy HPC code, which is usually relatively large. Hence, compiler techniques must be invented to do the translation.
- Transforming and scheduling our medium-grain dataflow model to meet some classic optimization criteria, such as throughput, local memory requirements, or I/O traffic.
- Combining agents and polyhedral kernels in dataflow languages. We propose to apply the techniques above to optimize the processes in actor-based dataflow languages and combine them with the parallelism existing in the languages.

We plan to rely extensively on the polyhedral model to define our compiler analysis. The polyhedral model was originally designed to analyze imperative programs. Analysis (such as scheduling or buffer allocation) must be redefined in light of dataflow semantics.

Translating a sequential program into a medium-grain dataflow model. The programs considered are compute-intensive parts from HPC applications, typically big HPC kernels of several hundreds of lines of C code. In particular, we expect to analyze the process code (actors) from the dataflow programs. On short ACL (Affine Control Loop) programs, direct solutions exist [56] and rely directly on array dataflow analysis [35]. On bigger ACL programs, this analysis no longer scales. We plan to address this issue by *modularizing* array dataflow analysis. Indeed, by splitting the program into processes, the complexity is mechanically reduced. This is a general observation, which was exploited in the past to compute schedules [34]. When the program is no longer ACL, a clear distinction must be made between polyhedral parts and non polyhedral parts. Hence, our medium-grain dataflow language must distinguish between polyhedral process networks, and non-polyhedral code fragments. This structure raises new challenges: How to abstract away non-polyhedral parts while keeping the polyhedrality of the dataflow program? Which trade-off(s) between precision and scalability are effective?

Medium-grain data transfers minimization. When the system consists of a single computing unit connected to a slow memory, the roofline model [59] defines the optimal ratio of computation per data transfer (*operational intensity*). The operational intensity is then translated to a partition of the computation (loop tiling) into *reuse units*: inside a reuse unit, data are transfered locally; between reuse units, data are transfered through the slow memory. On a *fine-grain* dataflow model, reuse units are exposed with loop tiling; this is the case for example

in Data-aware Process Network (DPN) [23]. The following questions are however still open: How does that translate on *medium-grain* dataflow models? And fundamentally what does it mean to *tile* a dataflow model?

Combining agents and polyhedral kernels in dataflow languages. In addition to the approach developed above, we propose to explore the compilation of dataflow programming languages. In fact, among the applications targeted by the project, some of them are already thought or specified as dataflow actors (video compression, machine-learning algorithms,...).

So far, parallelization techniques for such applications have focused on taking advantage of the decomposition into agents, potentially duplicating some agents to have several instances that work on different data items in parallel [40]. In the presence of big agents, the programmer is left with the splitting (or merging) of these agents by-hand if she wants to further parallelize her program (or at least give this opportunity to the runtime, which in general only sees agents as non-malleable entities). In the presence of arrays and loop-nests, or, more generally, some kind of regularity in the agent's code, however, we believe that the programmer would benefit from automatic parallelization techniques such as those proposed in the previous paragraphs. To achieve the goal of a totally integrated approach where programmers write the applications they have in mind (application flow in agents where the agents' code express potential parallelism), and then it is up to the system (compiler, runtime) to propose adequate optimizations, we propose to build on solid formal definition of the language semantics (thus the formal specification of parallelism occurring at the agent level) to provide hierarchical solutions to the problem of compilation and scheduling of such applications.

3.3.1. Expected impact

In general, splitting a program into simpler processes simplifies the problem. This observation leads to the following points:

- By abstracting away irregular parts in processes, we expect to structure the long-term problem of handling irregular applications in the polyhedral model. The long-term impact is to widen the applicability of the polyhedral model to irregular kernels.
- Splitting a program into processes reduces the problem size. Hence, it becomes possible to scale traditionally expensive polyhedral analysis such as scheduling or tiling to quote a few.

As for the third research direction, the short term impact is the possibility to combine efficiently classical dataflow programming with compiler polyhedral-based optimizations. We will first propose ad-hoc solutions coming from our HPC application expertise, but supported by strong theoretical results that prove their correctness and their applicability in practice. In the longer term, our work will allow specifying, designing, analyzing, and compiling HPC dataflow applications in a unified way. We target semi-automatic approaches where pertinent feedback is given to the developer during the development process.

3.3.2. Scientific Program

3.3.2.1. Short-term and ongoing activities.

We are currently working on the RTM (Reverse-Time Migration) kernel for oil and gas applications ($\approx$ 500 lines of C code). This kernel is long enough to be a good starting point, and small enough to be handled by a polyhedral splitting algorithm. We figured out the possible splittings so the polyhedral analysis can scale and irregular parts can be hidden. In a first step, we plan to define splitting metrics and algorithms to optimize the usual criteria: communication volume, latency and throughput.

Together with Lionel Morel (INSA/CEA LETI), we currently work on the evaluation of the practical advantage of combining the dataflow paradigm with the polyhedral optimization framework. We empirically build a proof-of-concept tooling approach, using existing tools on existing languages [38]. We combine dataflow programming with polyhedral compilation in order to enhance program parallelization by leveraging both inter-agent parallelism and intra-agent parallelism (i.e., regarding loop nests inside agents). We evaluate the approach practically, on benchmarks coming from image transformation or neural networks, and the first results demonstrate that there is indeed a room for further improvement.

3.3.2.2. Medium-term activities.

The results of the preceding paragraph are partial and have been obtained with a simple experimental approach only using off-the-shelf tools. We are thus encouraged to pursue research on combining expertise from dataflow programming languages and polyhedral compilation. Our long term objective is to go towards a formal framework to express, compile, and run dataflow applications with intrinsic instruction or pipeline parallelism.

We plan to investigate in the following directions:

- Investigate how polyhedral analysis extends on modular dataflow programs. For instance, how to modularize polyhedral scheduling analysis on our dataflow programs?
- Develop a proof of concept and validate it on linear algebra kernels (SVD, Gram-Schmidt, etc.).
- Explore various areas of applications from classical dataflow examples, like radio and video processing, to more recent applications in deep learning algorithmic. This will enable us to identify some potential (intra and extra) agent optimization patterns that could be leveraged into new language idioms.

3.3.2.3. Long-term activities.

Current work focus on purely polyhedral applications. Irregular parts are not handled. Also, a notion of tiling is required so the communications of the dataflow program with the outside world can be tuned with respect to the local memory size. Hence, we plan to investigate the following points:

- Assess simple polyhedral/non polyhedral partitioning: How non-polyhedral parts can be hidden in processes/channels? How to abstract the dataflow dependencies between processes? What would be the impact on analyses? We target programs with irregular control (e.g., while loop, early exits) and regular data (arrays with affine accesses).
- Design tiling schemes for modular dataflow programs: What does it mean to tile a dataflow program? Which compiler algorithms to use?
- Implement a mature compiler infrastructure from the front-end to code generation for a reasonable subset of the representation.

3.4. HLS-specific Dataflow Optimizations

The compiler analyses proposed in section 3.3 do not target a specific platform. In this part, we propose to leverage these analysis to develop source-level optimizations for high-level synthesis (HLS).

High-level synthesis consists in compiling a kernel written in a high-level language (typically in C) into a circuit. As for any compiler, an HLS tool consists in a *front-end* which translates the input kernel into an *intermediate representation*. This intermediate representation captures the control/flow dependences between computation units, generally in a hierarchical fashion. Then, the *back-end* maps this intermediate representation to a circuit (e.g. FPGA configuration). We believe that HLS tools must be thought as fine-grain automatic parallelizers. In classic HLS tools, the parallelism is expressed and exploited at the back-end level during the scheduling and the resource allocation of arithmetic operations. We believe that it would be far more profitable to derive the parallelism at the front-end level.

Hence, CASH will focus on the *front-end* pass and the *intermediate representation*. Low-level *back-end* techniques are not in the scope of CASH. Specifically, CASH will leverage the dataflow representation developed in Section 3.1 and the compilation techniques developed in Section 3.3 to develop a relevant intermediate representation for HLS and the corresponding front-end compilation algorithms.

Our results will be evaluated by using existing HLS tools (e.g., Intel HLS compiler, Xilinx Vivado HLS). We will implement our compiler as a source-to-source transformation in front of HLS tools. With this approach, HLS tools are considered as a "back-end black box". The CASH scheme is thus: (i) *front-end*: produce the CASH dataflow representation from the input C kernel. Then, (ii) turn this dataflow representation to a C program with pragmas for an HLS tool. This step must convey the characteristics of the dataflow representation found by step (i) (e.g. dataflow execution, fifo synchronisation, channel size). This source-to-source approach will allow us to get a full source-to-FPGA flow demonstrating the benefits of our tools while relying on existing tools for low-level optimizations. Step (i) will start from the DCC tool developed by Christophe Alias, which already produces a dataflow intermediate representation: the Data-aware Process Networks (DPN) [23]. Hence, the very first step is then to chose an HLS tool and to investiguate which input should be fed to the HLS tool so it "respects" the parallelism and the resource allocation suggested by the DPN. From this basis, we plan to investiguate the points described thereafter.

Roofline model and dataflow-level resource evaluation. Operational intensity must be tuned according to the roofline model. The roofline model [59] must be redefined in light of FPGA constraints. Indeed, the peak performance is no longer constant: it depends on the operational intensity itself. The more operational intensity we need, the more local memory we use, the less parallelization we get (since FPGA resources are limited), and finally the less performance we get! Hence, multiple iterations may be needed before reaching an efficient implementation. To accelerate the design process, we propose to iterate at the dataflow program level, which implies a fast resource evaluation at the dataflow level.

Reducing FPGA resources. Each parallel unit must use as little resources as possible to maximize parallel duplication, hence the final performance. This requires to factorize the control and the channels. Both can be achieved with source-to-source optimizations at dataflow level. The main issue with outputs from polyhedral optimization is large piecewise affine functions that require a wide silicon surface on the FPGA to be computed. Actually we do not need to compute a closed form (expression that can be evaluated in bounded time on the FPGA) *statically*. We believe that the circuit can be compacted if we allow control parts to be evaluated dynamically. Finally, though dataflow architectures are a natural candidate, adjustments are required to fit FPGA constraints (2D circuit, few memory blocks). Ideas from systolic arrays [52] can be borrowed to re-use the same piece of data multiple times, despite the limitation to regular kernels and the lack of I/O flexibility. A trade-off must be found between pure dataflow and systolic communications.

Improving circuit throughput. Since we target streaming applications, the throughput must be optimized. To achieve such an optimization, we need to address the following questions. How to derive an optimal upper bound on the throughput for polyhedral process network? Which dataflow transformations should be performed to reach it? The limiting factors are well known: I/O (decoding of burst data), communications through addressable channels, and latencies of the arithmetic operators. Finally, it is also necessary to find the right methodology to measure the throughput statically and/or dynamically.

3.4.1. Expected Impact

So far, the HLS front-end applies basic loop optimizations (unrolling, flattening, pipelining, etc.) and use a Hierarchical Control Flow Graph-like representation with data dependencies annotations (HCDFG). With this approach, we intend to demonstrate that polyhedral analysis combined with dataflow representations is an effective solution for HLS tools.

3.4.2. Scientific Program

3.4.2.1. Short-term and ongoing activities.

The HLS compiler designed in the CASH team currently extracts a fine-grain parallel intermediate representation (DPN [23], [22]) from a sequential program. We will not write a back-end that produces code for FPGA but we need to provide C programs that can be fed into existing C-to-FPGA compilers. However we obviously need an end-to-end compiler for our experiments. One of the first task of our HLS activity is to develop a DPN-to-C code generator suitable as input to an existing HLS tool like Vivado HLS. The generated code should exhibit the parallelism extracted by our compiler, and allow generating a final circuit more efficient than

the one that would be generated by our target HLS tool if ran directly on the input program. Source-to-source approaches have already been experimented successfully, e.g. in Alexandru Plesco's PhD [51].

3.4.2.2. Medium-term activities.

Our DPN-to-C code generation will need to be improved in many directions. The first point is the elimination of redundancies induced by the DPN model itself: buffers are duplicated to allow parallel reads, processes are produced from statements in the same loop, hence with the same control automaton. Also, multiplexing uses affine constraints which can be factorized [24]. We plan to study how these constructs can be factorized at C-level and to design the appropriate DPN-to-C translation algorithms.

Also, we plan to explore how on-the-fly evaluation can reduce the complexity of the control. A good starting point is the control required for the load process (which fetch data from the distant memory). If we want to avoid multiple load of the same data, the FSM (Finite State Machine) that describes it is usually very complex. We believe that dynamic construction of the load set (set of data to load from the main memory) will use less silicon than an FSM with large piecewise affine functions computed statically.

3.4.2.3. Long-term activities.

The DPN-to-C compiler opens new research perspectives. We will explore the roofline model accuracy for different applications by playing on DPN parameters (tile size). Unlike the classical roofline model, the peak performance is no longer assumed to be constant, but decreasing with operational intensity [60]. Hence, we expect a *unique* optimal set of parameters. Thus, we need to build a DPN-level cost model to derive an interval containing the optimal parameters.

Also, we want to develop DPN-level analysis and transformation to quantify the optimal reachable throughput and to reach it. We expect the parallelism to increase the throughput, but in turn it may require an operational intensity beyond the optimal point discussed in the first paragraph. We will assess the trade-offs, build the cost-models, and the relevant dataflow transformations.

3.5. Simulation of Hardware

Complex systems such as systems-on-a-chip or HPC computer with FPGA accelerator comprise both hardware and software parts, tightly coupled together. In particular, the software cannot be executed without the hardware, or at least a simulator of the hardware.

Because of the increasing complexity of both software and hardware, traditional simulation techniques (Register Transfer Level, RTL) are too slow to allow full system simulation in reasonable time. New techniques such as Transaction Level Modeling (TLM) [20] in SystemC [19] have been introduced and widely adopted in the industry. Internally, SystemC uses discrete-event simulation, with efficient context-switch using cooperative scheduling. TLM abstracts away communication details, and allows modules to communicate using function calls. We are particularly interested in the loosely timed coding style where the timing of the platform is not modeled precisely, and which allows the fastest simulations. This allowed gaining several orders of magnitude of simulation speed. However, SystemC/TLM is also reaching its limits in terms of performance, in particular due to its lack of parallelism.

Work on SystemC/TLM parallel execution is both an application of other work on parallelism in the team and a tool complementary to HLS presented in Sections 3.1 (dataflow models and programs) and 3.4 (application to FPGA). Indeed, some of the parallelization techniques we develop in CASH could apply to SystemC/TLM programs. Conversely, a complete design-flow based on HLS needs fast system-level simulation: the full-system usually contains both hardware parts designed using HLS, handwritten hardware components, and software.

We also work on simulation of the DPN intermediate representation. Simulation is a very important tool to help validate and debug a complete compiler chain. Without simulation, validating the front-end of the compiler requires running the full back-end and checking the generated circuit. Simulation can avoid the execution time of the backend and provide better debugging tools.

Automatic parallelization has shown to be hard, if at all possible, on loosely timed models [28]. We focus on semi-automatic approaches where the programmer only needs to make minor modifications of programs to get significant speedups.

3.5.1. Expected Impact

The short term impact is the possibility to improve simulation speed with a reasonable additional programming effort. The amount of additional programming effort will thus be evaluated in the short term.

In the longer term, our work will allow scaling up simulations both in terms of models and execution platforms. Models are needed not only for individual Systems on a Chip, but also for sets of systems communicating together (e.g., the full model for a car which comprises several systems communicating together), and/or heterogeneous models. In terms of execution platform, we are studying both parallel and distributed simulations.

3.5.2. Scientific Program

3.5.2.1. Short-term and ongoing activities.

We started the joint PhD (with Tanguy Sassolas) of Gabriel Busnot with CEA-LIST. The research targets parallelizing SystemC heterogeneous simulations. CEA-LIST already developed SCale [57], which is very efficient to simulate parallel homogeneous platforms such as multi-core chips. However, SCale cannot currently load-balance properly the computations when the platform contains different components modeled at various levels of abstraction. Also, SCale requires manual annotations to identify accesses to shared variables. These annotations are given as address ranges in the case of a shared memory. This annotation scheme does not work when the software does non-trivial memory management (virtual memory using a memory management unit, dynamic allocation), since the address ranges cannot be known statically. We started working on the "heterogeneous" aspect of simulations with an approach allowing changing the level of details in a simulation at runtime, and started tackling the virtual and dynamic memory management problem by porting Linux on our simulation platform.

We also started working on an improved support for simulation and debugging of the DPN internal representation of our parallelizing compiler (see Section 3.3). A previous quick experiment with simulation was to generate C code that simulates parallelism with POSIX-threads. While this simulator greatly helped debug the compiler, this is limited in several ways: simulations are not deterministic, and the simulator does not scale up since it would create a very large number of threads for a non-trivial design.

We are working in two directions. The first is to provide user-friendly tools to allow graphical inspection of traces. For example, we started working on the visualization of the sequence of steps leading to a deadlock when the situation occurs, and will give hints on how to fix the problem in the compiler. The second is to use an efficient simulator to speed up the simulation. We plan to generate SystemC/TLM code from the DPN representation to benefit from the ability of SystemC to simulate a large number of processes.

3.5.2.2. Medium-term activities.

Several research teams have proposed different approaches to deal with parallelism and heterogeneity. Each approach targets a specific abstraction level and coding style. While we do not hope for a universal solution, we believe that a better coordination of different actors of the domain could lead to a better integration of solutions. We could imagine, for example, a platform with one subsystem accelerated with SCale [57] from CEA-LIST, some compute-intensive parts delegated to sc-during [49] from Matthieu Moy, and a co-simulation with external physical solvers using SystemC-MDVP [27] from LIP6. We plan to work on the convergence of approaches, ideally both through point-to-point collaborations and with a collaborative project.

A common issue with heterogeneous simulation is the level of abstraction. Physical models only simulate one scenario and require concrete input values, while TLM models are usually abstract and not aware of precise physical values. One option we would like to investigate is a way to deal with loose information, e.g. manipulate intervals of possible values instead of individual, concrete values. This would allow a simulation to be symbolic with respect to the physical values.

Obviously, works on parallel execution of simulations would benefit to simulation of data-aware process networks (DPN). Since DPN are generated, we can even tweak the generator to guarantee some properties on the generated code, which gives us more freedom on the parallelization and partitioning techniques.

3.5.2.3. Long-term activities.

In the long term, our vision is a simulation framework that will allow combining several simulators (not necessarily all SystemC-based), and allow running them in a parallel way. The Functional Mockup Interface (FMI) standard is a good basis to build upon, but the standard does not allow expressing timing and functional constraints needed for a full co-simulation to run properly.

CELTIQUE Project-Team (section vide)

CIDRE Project-Team

3. Research Program

3.1. Our perspective

For many aspects of our daily lives, we rely heavily on computer systems, many of which are based on massively interconnected devices that support a population of interacting and cooperating entities. As these systems become more open and complex, accidental and intentional failures become much more frequent and serious. We believe that the purpose of attacks against these systems is expressed at a high level (compromise of sensitive data, unavailability of services). However, these attacks are often carried out at a very low level (exploitation of vulnerabilities by malicious code, hardware attacks).

The CIDRE team is specialized in the defense of computer systems. We argue that to properly protect these systems we must have a complete understanding of the attacker's concrete capabilities. In other words, **to defend properly we must understand the attack**.

The CIDRE team therefore strives to have a global expertise in information systems: from hardware to distributed architectures. Our objective is to highlight security issues and propose preventive or reactive countermeasures in widely used and privacy-friendly systems.

3.2. Attack Comprehension

An attack on a computer system begins with the exploitation of one or more vulnerabilities of that system. Generally speaking, a vulnerability can be a software bug or a misconfiguration that can be exploited by the attacker to perform unauthorized actions. Exploiting a vulnerability leads to a use of the system according to a case not foreseen in its specification, implementation or configuration. This puts the system in an inconsistent state allowing the attacker to divert the use of the system in his or her own interest.

The systems we use are large, interconnected, constantly evolving and, therefore, are likely to retain many vulnerabilities; their security depends on our ability to update them quickly when new threats are discovered. It is thus necessary to understand how the attacker has compromised the system: what vulnerabilities he has exploited, what actions he has conducted, where he is located in the system. It is essential to study statically the malicious code used by the attacker. It is also important to be able to study it dynamically to be able to replay attacks on demand.

Ideally, we should be ahead of the attacker and therefore imagine new ways to attack. In addition, we believe it is necessary to improve the feedback to the expert by allowing him to quickly understand the progress of an attack. The first step before being able to offer secure systems is to understand and measure the real capabilities of the attacker.

Our first research axis therefore aims at highlighting both the effective attacker's means and the way an attack unfolds and spreads.

In this context, we are particularly interested in

- **highlighting attacks** on the micro-architecture that affect software security
- **providing expert support**
  - to analyze malicious code
  - to quickly investigate an intrusion on a system monitored by an intrusion detection system

3.3. Attack Detection

An attack is generally composed of several steps. During a first approach step the attacker enters the system, locates the target and makes itself persistent. Then, in a second step, the payload of the attack is effectively launched, leading to a violation of the security policy (attacks against confidentiality, integrity, or availability of OS, applications, services, or data).

The objective of intrusion detection is to be able to detect the attacker, ideally during the first step of the attack. To do this, intrusion detection systems (IDS) are based on probes that continuously monitor the system. These probes report events to a core engine that decide whether or not to alert the expert.

Intrusion detection systems are important for all systems handling sensitive data that may be accessible to a malicious agent. They are especially crucial for low-level systems that provide essential support services to other systems. They are essential in inter-connected systems that are designed to last a long time and are difficult to update.

3.4. Attack Resistance

The first two axes of the team allowed us to measure the concrete technical means of the attacker. We claim that the attacker can always avoid the measures put in place to secure a system. We believe that another way to offer more secure systems is to take into account from the design phase that these systems will operate in the presence of an omnipotent attacker. The last research axis of the CIDRE team is focused on offering systems that are resistant to attackers, *i.e.* they can provide the expected services even in the presence of an attacker.

To achieve this goal, we explore two approaches:

- deceptive security
- malicious behavior tolerance

In the notion of *deceptive security* we group together all the approaches that aim to mislead the active attacker in a system in order to deceive him on the exact nature of his target. These approaches can slow down the attacker or lead him to abandon his attack.

Finally, we contribute to the design of architectures or services relying on the collaboration of entities that is not affected by the minority presence of malicious entities. These architectures or services are based on the collaboration of a set of nodes that are not affected by the presence in minority of malicious nodes.

COMETE Project-Team

3. Research Program

3.1. Probability and information theory

Participants: Konstantinos Chatzikokolakis, Catuscia Palamidessi, Marco Romanelli, Anna Pazii.

Much of the research of Comète focuses on security and privacy. In particular, we are interested in the problem of the leakage of secret information through public observables.

Ideally we would like systems to be completely secure, but in practice this goal is often impossible to achieve. Therefore, we need to reason about the amount of information leaked, and the utility that it can have for the adversary, i.e. the probability that the adversary is able to exploit such information.

The recent tendency is to use an information theoretic approach to model the problem and define the leakage in a quantitative way. The idea is to consider the system as an information-theoretic *channel*. The input represents the secret, the output represents the observable, and the correlation between the input and output (*mutual information*) represents the information leakage.

Information theory depends on the notion of entropy as a measure of uncertainty. From the security point of view, this measure corresponds to a particular model of attack and a particular way of estimating the security threat (vulnerability of the secret). Most of the proposals in the literature use Shannon entropy, which is the most established notion of entropy in information theory. We, however, consider also other notions, in particular Rényi min-entropy, which seems to be more appropriate for security in common scenarios like one-try attacks.

3.2. Expressiveness of Concurrent Formalisms

Participants: Catuscia Palamidessi, Frank Valencia.

We study computational models and languages for distributed, probabilistic and mobile systems, with a particular attention to expressiveness issues. We aim at developing criteria to assess the expressive power of a model or formalism in a distributed setting, to compare existing models and formalisms, and to define new ones according to an intended level of expressiveness, also taking into account the issue of (efficient) implementability.

3.3. Concurrent constraint programming

Participants: Frank Valencia, Santiago Quintero.

Concurrent constraint programming (ccp) is a well established process calculus for modeling systems where agents interact by posting and asking information in a store, much like in users interact in *social networks*. This information is represented as first-order logic formulae, called constraints, on the shared variables of the system (e.g., $X > 42$). The most distinctive and appealing feature of ccp is perhaps that it unifies in a single formalism the operational view of processes based upon process calculi with a declarative one based upon first-order logic. It also has an elegant denotational semantics that interprets processes as closure operators (over the set of constraints ordered by entailment). In other words, any ccp process can be seen as an idempotent, increasing, and monotonic function from stores to stores. Consequently, ccp processes can be viewed as: computing agents, formulae in the underlying logic, and closure operators. This allows ccp to benefit from the large body of techniques of process calculi, logic and domain theory.

Our research in ccp develops along the following two lines:

1. **(a)** The study of a bisimulation semantics for ccp. The advantage of bisimulation, over other kinds of semantics, is that it can be efficiently verified.
2. **(b)** The extension of ccp with constructs to capture emergent systems such as those in social networks and cloud computing.

3.4. Model checking

Participants: Konstantinos Chatzikokolakis, Catuscia Palamidessi.

Model checking addresses the problem of establishing whether a given specification satisfies a certain property. We are interested in developing model-checking techniques for verifying concurrent systems of the kind explained above. In particular, we focus on security and privacy, i.e., on the problem of proving that a given system satisfies the intended security or privacy properties. Since the properties we are interested in have a probabilistic nature, we use probabilistic automata to model the protocols. A challenging problem is represented by the fact that the interplay between nondeterminism and probability, which in security presents subtleties that cannot be handled with the traditional notion of a scheduler,

CONVECS Project-Team

3. Research Program

3.1. New Formal Languages and their Concurrent Implementations

We aim at proposing and implementing new formal languages for the specification, implementation, and verification of concurrent systems. In order to provide a complete, coherent methodological framework, two research directions must be addressed:

- *Model-based specifications*: these are operational (i.e., constructive) descriptions of systems, usually expressed in terms of processes that execute concurrently, synchronize together and communicate. Process calculi are typical examples of model-based specification languages. The approach we promote is based on LOTOS NT (LNT for short), a formal specification language that incorporates most constructs stemming from classical programming languages, which eases its acceptance by students and industry engineers. LNT [6] is derived from the ISO standard E-LOTOS (2001), of which it represents the first successful implementation, based on a source-level translation from LNT to the former ISO standard LOTOS (1989). We are working both on the semantic foundations of LNT (enhancing the language with module interfaces and timed/probabilistic/stochastic features, compiling the m among n synchronization, etc.) and on the generation of efficient parallel and distributed code. Once equipped with these features, LNT will enable formally verified asynchronous concurrent designs to be implemented automatically.
- *Property-based specifications*: these are declarative (i.e., non-constructive) descriptions of systems, which express *what* a system should do rather than *how* the system should do it. Temporal logics and μ-calculi are typical examples of property-based specification languages. The natural models underlying value-passing specification languages, such as LNT, are Labeled Transition Systems (LTSs or simply *graphs*) in which the transitions between states are labeled by actions containing data values exchanged during handshake communications. In order to reason accurately about these LTSs, temporal logics involving data values are necessary. The approach we promote is based on MCL (*Model Checking Language*) [47], which extends the modal μ-calculus with data-handling primitives, fairness operators encoding generalized Büchi automata, and a functional-like language for describing complex transition sequences. We are working both on the semantic foundations of MCL (extending the language with new temporal and hybrid operators, translating these operators into lower-level formalisms, enhancing the type system, etc.) and also on improving the MCL on-the-fly model checking technology (devising new algorithms, enhancing ergonomy by detecting and reporting vacuity, etc.).

We address these two directions simultaneously, yet in a coherent manner, with a particular focus on applicable concurrent code generation and computer-aided verification.

3.2. Parallel and Distributed Verification

Exploiting large-scale high-performance computers is a promising way to augment the capabilities of formal verification. The underlying problems are far from trivial, making the correct design, implementation, fine-tuning, and benchmarking of parallel and distributed verification algorithms long-term and difficult activities. Sequential verification algorithms cannot be reused as such for this task: they are inherently complex, and their existing implementations reflect several years of optimizations and enhancements. To obtain good speedup and scalability, it is necessary to invent new parallel and distributed algorithms rather than to attempt a parallelization of existing sequential ones. We seek to achieve this objective by working along two directions:

- *Rigorous design:* Because of their high complexity, concurrent verification algorithms should themselves be subject to formal modeling and verification, as confirmed by recent trends in the certification of safety-critical applications. To facilitate the development of new parallel and distributed verification algorithms, we promote a rigorous approach based on formal methods and verification. Such algorithms will be first specified formally in LNT, then validated using existing model checking algorithms of the CADP toolbox. Second, parallel or distributed implementations of these algorithms will be generated automatically from the LNT specifications, enabling them to be experimented on large computing infrastructures, such as clusters and grids. As a side-effect, this "bootstrapping" approach would produce new verification tools that can later be used to self-verify their own design.
- *Performance optimization:* In devising parallel and distributed verification algorithms, particular care must be taken to optimize performance. These algorithms will face concurrency issues at several levels: grids of heterogeneous clusters (architecture-independence of data, dynamic load balancing), clusters of homogeneous machines connected by a network (message-passing communication, detection of stable states), and multi-core machines (shared-memory communication, thread synchronization). We will seek to exploit the results achieved in the parallel and distributed computing field to improve performance when using thousands of machines by reducing the number of connections and the messages exchanged between the cooperating processes carrying out the verification task. Another important issue is the generalization of existing LTS representations (explicit, implicit, distributed) in order to make them fully interoperable, such that compilers and verification tools can handle these models transparently.

3.3. Timed, Probabilistic, and Stochastic Extensions

Concurrent systems can be analyzed from a *qualitative* point of view, to check whether certain properties of interest (e.g., safety, liveness, fairness, etc.) are satisfied. This is the role of functional verification, which produces Boolean (yes/no) verdicts. However, it is often useful to analyze such systems from a *quantitative* point of view, to answer non-functional questions regarding performance over the long run, response time, throughput, latency, failure probability, etc. Such questions, which call for numerical (rather than binary) answers, are essential when studying the performance and dependability (e.g., availability, reliability, etc.) of complex systems.

Traditionally, qualitative and quantitative analyzes are performed separately, using different modeling languages and different software tools, often by distinct persons. Unifying these separate processes to form a seamless design flow with common modeling languages and analysis tools is therefore desirable, for both scientific and economic reasons. Technically, the existing modeling languages for concurrent systems need to be enriched with new features for describing quantitative aspects, such as probabilities, weights, and time. Such extensions have been well-studied and, for each of these directions, there exist various kinds of automata, e.g., discrete-time Markov chains for probabilities, weighted automata for weights, timed automata for hard real-time, continuous-time Markov chains for soft real-time with exponential distributions, etc. Nowadays, the next scientific challenge is to combine these individual extensions altogether to provide even more expressive models suitable for advanced applications.

Many such combinations have been proposed in the literature, and there is a large amount of models adding probabilities, weights, and/or time. However, an unfortunate consequence of this diversity is the confuse landscape of software tools supporting such models. Dozens of tools have been developed to implement theoretical ideas about probabilities, weights, and time in concurrent systems. Unfortunately, these tools do not interoperate smoothly, due both to incompatibilities in the underlying semantic models and to the lack of common exchange formats.

To address these issues, CONVECS follows two research directions:

- *Unifying the semantic models*. Firstly, we will perform a systematic survey of the existing semantic models in order to distinguish between their essential and non-essential characteristics, the goal being to propose a unified semantic model that is compatible with process calculi techniques for specifying and verifying concurrent systems. There are already proposals for unification either

theoretical (e.g., Markov automata) or practical (e.g., PRISM and MODEST modeling languages), but these languages focus on quantitative aspects and do not provide high-level control structures and data handling features (as LNT does, for instance). Work is therefore needed to unify process calculi and quantitative models, still retaining the benefits of both worlds.

- *Increasing the interoperability of analysis tools*. Secondly, we will seek to enhance the interoperability of existing tools for timed, probabilistic, and stochastic systems. Based on scientific exchanges with developers of advanced tools for quantitative analysis, we plan to evolve the CADP toolbox as follows: extending its perimeter of functional verification with quantitative aspects; enabling deeper connections with external analysis components for probabilistic, stochastic, and timed models; and introducing architectural principles for the design and integration of future tools, our long-term goal being the construction of a European collaborative platform encompassing both functional and non-functional analyzes.

3.4. Component-Based Architectures for On-the-Fly Verification

On-the-fly verification fights against state explosion by enabling an incremental, demand-driven exploration of LTSs, thus avoiding their entire construction prior to verification. In this approach, LTS models are handled implicitly by means of their *post* function, which computes the transitions going out of given states and thus serves as a basis for any forward exploration algorithm. On-the-fly verification tools are complex software artifacts, which must be designed as modularly as possible to enhance their robustness, reduce their development effort, and facilitate their evolution. To achieve such a modular framework, we undertake research in several directions:

- *New interfaces for on-the-fly LTS manipulation*. The current application programming interface (API) for on-the-fly graph manipulation, named OPEN/CAESAR [35], provides an "opaque" representation of states and actions (transitions labels): states are represented as memory areas of fixed size and actions are character strings. Although appropriate to the pure process algebraic setting, this representation must be generalized to provide additional information supporting an efficient construction of advanced verification features, such as: handling of the types, functions, data values, and parallel structure of the source program under verification, independence of transitions in the LTS, quantitative (timed/probabilistic/stochastic) information, etc.
- *Compositional framework for on-the-fly LTS analysis*. On-the-fly model checkers and equivalence checkers usually perform several operations on graph models (LTSs, Boolean graphs, etc.), such as exploration, parallel composition, partial order reduction, encoding of model checking and equivalence checking in terms of Boolean equation systems, resolution and diagnostic generation for Boolean equation systems, etc. To facilitate the design, implementation, and usage of these functionalities, it is necessary to encapsulate them in software components that could be freely combined and replaced. Such components would act as graph transformers, that would execute (on a sequential machine) in a way similar to coroutines and to the composition of lazy functions in functional programming languages. Besides its obvious benefits in modularity, such a component-based architecture will also make it possible to take advantage of multi-core processors.
- *New generic components for on-the-fly verification*. The quest for new on-the-fly components for LTS analysis must be pursued, with the goal of obtaining a rich catalog of interoperable components serving as building blocks for new analysis features. A long-term goal of this approach is to provide an increasingly large catalog of interoperable components covering all verification and analysis functionalities that appear to be useful in practice. It is worth noticing that some components can be very complex pieces of software (e.g., the encapsulation of an on-the-fly model checker for a rich temporal logic). Ideally, it should be possible to build a novel verification or analysis tool by assembling on-the-fly graph manipulation components taken from the catalog. This would provide a flexible means of building new verification and analysis tools by reusing generic, interoperable model manipulation components.

3.5. Real-Life Applications and Case Studies

We believe that theoretical studies and tool developments must be confronted with significant case studies to assess their applicability and to identify new research directions. Therefore, we seek to apply our languages, models, and tools for specifying and verifying formally real-life applications, often in the context of industrial collaborations.

CORSE Project-Team

3. Research Program

3.1. Scientific Foundations

One of the characteristics of CORSE is to base our researches on diverse advanced mathematical tools. Compiler optimization requires the usage of the several tools around discrete mathematics: combinatorial optimization, algorithmic, and graph theory. The aim of CORSE is to tackle optimization not only for general purpose but also for domain specific applications. We believe that new challenges in compiler technology design and in particular for split compilation should also take advantage of graph labeling techniques. In addition to run-time and compiler techniques for program instrumentation, hybrid analysis and compilation advances will be mainly based on polynomial and linear algebra.

The other specificity of CORSE is to address technical challenges related to compiler technology, run-time systems, and hardware characteristics. This implies mastering the details of each. This is especially important as any optimization is based on a reasonably accurate model. Compiler expertise will be used in modeling applications (e.g. through automatic analysis of memory and computational complexity); Run-time expertise will be used in modeling the concurrent activities and overhead due to contention (including memory management); Hardware expertise will be extensively used in modeling physical resources and hardware mechanisms (including synchronization, pipelines, etc.).

The core foundation of the team is related to the combination of static and dynamic techniques, of compilation, and run-time systems. We believe this to be essential in addressing high-performance and low energy challenges in the context of new important changes shown by current application, software, and architecture trends.

Our project is structured along two main directions. The first direction belongs to the area of run-time systems with the objective of developing strong relations with compilers. The second direction belongs to the area of compiler analysis and optimization with the objective of combining dynamic analysis and optimization with static techniques. The aim of CORSE is to ground those two research activities on the development of the end-to-end optimization of some specific domain applications.

DATASHAPE Project-Team

3. Research Program

3.1. Algorithmic aspects of topological and geometric data analysis

TDA requires to construct and manipulate appropriate representations of complex and high dimensional shapes. A major difficulty comes from the fact that the complexity of data structures and algorithms used to approximate shapes rapidly grows as the dimensionality increases, which makes them intractable in high dimensions. We focus our research on simplicial complexes which offer a convenient representation of general shapes and generalize graphs and triangulations. Our work includes the study of simplicial complexes with good approximation properties and the design of compact data structures to represent them.

In low dimensions, effective shape reconstruction techniques exist that can provide precise geometric approximations very efficiently and under reasonable sampling conditions. Extending those techniques to higher dimensions as is required in the context of TDA is problematic since almost all methods in low dimensions rely on the computation of a subdivision of the ambient space. A direct extension of those methods would immediately lead to algorithms whose complexities depend exponentially on the ambient dimension, which is prohibitive in most applications. A first direction to by-pass the curse of dimensionality is to develop algorithms whose complexities depend on the intrinsic dimension of the data (which most of the time is small although unknown) rather than on the dimension of the ambient space. Another direction is to resort to cruder approximations that only captures the homotopy type or the homology of the sampled shape. The recent theory of persistent homology provides a powerful and robust tool to study the homology of sampled spaces in a stable way.

3.2. Statistical aspects of topological and geometric data analysis

The wide variety of larger and larger available data - often corrupted by noise and outliers - requires to consider the statistical properties of their topological and geometric features and to propose new relevant statistical models for their study.

There exist various statistical and machine learning methods intending to uncover the geometric structure of data. Beyond manifold learning and dimensionality reduction approaches that generally do not allow to assert the relevance of the inferred topological and geometric features and are not well-suited for the analysis of complex topological structures, set estimation methods intend to estimate, from random samples, a set around which the data is concentrated. In these methods, that include support and manifold estimation, principal curves/manifolds and their various generalizations to name a few, the estimation problems are usually considered under losses, such as Hausdorff distance or symmetric difference, that are not sensitive to the topology of the estimated sets, preventing these tools to directly infer topological or geometric information.

Regarding purely topological features, the statistical estimation of homology or homotopy type of compact subsets of Euclidean spaces, has only been considered recently, most of the time under the quite restrictive assumption that the data are randomly sampled from smooth manifolds.

In a more general setting, with the emergence of new geometric inference tools based on the study of distance functions and algebraic topology tools such as persistent homology, computational topology has recently seen an important development offering a new set of methods to infer relevant topological and geometric features of data sampled in general metric spaces. The use of these tools remains widely heuristic and until recently there were only a few preliminary results establishing connections between geometric inference, persistent homology and statistics. However, this direction has attracted a lot of attention over the last three years. In particular, stability properties and new representations of persistent homology information have led to very promising results to which the DATASHAPE members have significantly contributed. These preliminary results open many perspectives and research directions that need to be explored.

Our goal is to build on our first statistical results in TDA to develop the mathematical foundations of Statistical Topological and Geometric Data Analysis. Combined with the other objectives, our ultimate goal is to provide a well-founded and effective statistical toolbox for the understanding of topology and geometry of data.

3.3. Topological approach for multimodal data processing

Due to their geometric nature, multimodal data (images, video, 3D shapes, etc.) are of particular interest for the techniques we develop. Our goal is to establish a rigorous framework in which data having different representations can all be processed, mapped and exploited jointly. This requires adapting our tools and sometimes developing entirely new or specialized approaches.

The choice of multimedia data is motivated primarily by the fact that the amount of such data is steadily growing (with e.g. video streaming accounting for nearly two thirds of peak North-American Internet traffic, and almost half a billion images being posted on social networks each day), while at the same time it poses significant challenges in designing informative notions of (dis)-similarity as standard metrics (e.g. Euclidean distances between points) are not relevant.

3.4. Experimental research and software development

We develop a high quality open source software platform called GUDHI which is becoming a reference in geometric and topological data analysis in high dimensions. The goal is not to provide code tailored to the numerous potential applications but rather to provide the central data structures and algorithms that underlie applications in geometric and topological data analysis.

The development of the GUDHI platform also serves to benchmark and optimize new algorithmic solutions resulting from our theoretical work. Such development necessitates a whole line of research on software architecture and interface design, heuristics and fine-tuning optimization, robustness and arithmetic issues, and visualization. We aim at providing a full programming environment following the same recipes that made up the success story of the CGAL library, the reference library in computational geometry.

Some of the algorithms implemented on the platform will also be interfaced to other software platform, such as the R software [0] for statistical computing, and languages such as Python in order to make them usable in combination with other data analysis and machine learning tools. A first attempt in this direction has been done with the creation of an R package called TDA in collaboration with the group of Larry Wasserman at Carnegie Mellon University (Inria Associated team CATS) that already includes some functionalities of the GUDHI library and implements some joint results between our team and the CMU team. A similar interface with the Python language is also considered a priority. To go even further towards helping users, we will provide utilities that perform the most common tasks without requiring any programming at all.

[0] https://www.r-project.org/

DATASPHERE Team

3. Research Program

3.1. Dynamics of digital transformations

The research program of the Datasphere team aims at understanding the transformations induced by digital systems on socio-economic and socio-ecological organizations. These transformations are very broad and impact a large part of society. Understanding these changes is very ambitious and would require much more resources than those of the team. Interactions with other teams in other disciplines is thus of strategic importance. The research directions we have worked in and will continue to in the coming years are the following.

- The legal and strategic implications of the development of networks, the growing global interdependencies, and the increase of digital flows beyond control.
- The geopolitics of digital systems, data flows and cyber control, the raise of new strategic imbalances, and digital powers (US, China, Russia, etc.)
- The structural consequences of the translation of governance to digital actors, their inclusion into diplomatic forums, and the weakening of sovereignty over territories.

3.2. Foundations of digital economy

- The economy of intermediation and the progressive control of all two-sided and multi-sided markets by remote digital platforms.
- The methodologies for assessing the strategic value of data and evaluating its leverage for the political economy.
- The analysis of Online Advertisement/tracking ecosystems.

3.3. Ecosystems and Anthropocene

- The interdependencies of natural ecosystems and socio-economic systems, and the role of digital systems on measuring and controlling the global natural/social system.
- The role of digital actors in the adaptation and mitigation of climate change.
- The information economy of planetary challenges related to global warming, biodiversity, health monitoring.

3.4. Large scale graph analysis

- Community analysis and extraction, spectral methods.
- Manifold based approaches to large scale graph analysis, optimal transport.
- Information/rumor/fake news propagation in social networks.

3.5. Cyberstrategy

- Geopolitics of BGP
- Cyberstrategy of infrastructures
- Internet content control

DEDUCTEAM Project-Team

3. Research Program

3.1. Logical Frameworks

A thesis, which is at the root of our research effort, is that logical systems should be expressed as theories in a logical framework. As a consequence, proof-checking systems should not be focused on one theory, such as Simple type theory, Martin-Löf's type theory, or the Calculus of constructions, but should be theory independent. On the more theoretical side, the proof search algorithms, or the algorithmic interpretation of proofs should not depend on the theory in which proofs are expressed, but this theory should just be a parameter. This is for instance expressed in the title of our invited talk at ICALP 2012: *A theory independent Curry-De Bruijn-Howard correspondence* [25].

Various limits of Predicate logic have led to the development of various families of logical frameworks: λ-prolog and Isabelle have allowed terms containing free variables, the Edinburgh logical framework has allowed proofs to be expressed as λ-terms, Pure type systems have allowed propositions to be considered as terms, and Deduction modulo theory has allowed theories to be defined not only with axioms, but also with computation rules.

The $\lambda\Pi$-calculus modulo theory, that is implemented in the system DEDUKTI and that is a synthesis of the Edinburgh logical framework and of Deduction modulo theory, subsumes them all. Part of our research effort is focused on improving the $\lambda\Pi$-calculus modulo theory, for instance allowing to define congruences with associative and commutative rewriting. Another part of our research effort is focused on the automatic analysis of theories to prove their confluence, termination, and consistency either by pencil and paper proofs or automatically [4].

3.2. Interoperability and proof encyclopediae

Using a single prover to check proofs coming from different systems naturally leads to investigate how these proofs can be translated from one theory to another and used in a system different from the system in which they have been developed. This issue is of prime importance because developments in proof systems are getting bigger and, unlike other communities in computer science, the proof checking community has given little effort in the direction of standardization and interoperability.

For each proof, independently of the system in which it has been developed, we should be able to identify the systems in which it can be expressed. For instance, we have shown that many proofs developed in the MATITA prover did not use the full strength of the logic of MATITA and could be exported, for instance, to the systems of the HOL family, that are based on a weaker logic.

Rather than importing proofs from one system, transforming them, and exporting them to another system, we can use the same tools to develop system-independent proof encyclopedia called Logipedia. In such a library, each proof is labeled with the theories in which it can be expressed and so with the systems in which it can be used.

3.3. Interactive theorem proving

If our main goal with DEDUKTI is to import, transform, and export proofs developed in other systems, we also want to investigate how DEDUKTI can be used as the basis of an interactive theorem prover. This leads to two new scientific questions: first, how much can a tactic system be theory independent, and then how does rewriting extends the possibility to write tactics.

This has led to the development of a new version of DEDUKTI, which supports metavariables. Several tactics have been developed for this system, which are intended to help a human user to write proofs in our system instead of writing proof terms by hand. This work is a continuation of the previous work the team did on DEMON, which was an extension of DEDUKTI, whereas the support for interactive theorem proving is now native in DEDUKTI.

GALLINETTE Project-Team

3. Research Program

3.1. Scientific Context

Software quality is a requirement that is becoming more and more prevalent, by now far exceeding the traditional scope of embedded systems. The development of tools to construct software that respects a given specification is a major challenge facing computer science. *Proof assistants* such as Coq [49] provide a formal method whose central innovation is to produce *certified programs* by transforming the very activity of programming. Programming and proving are merged into a single development activity, informed by an elegant but rigid mathematical theory inspired by the correspondence between programming, logic and algebra: the *Curry-Howard correspondence*. For the certification of programs, this approach has shown its efficiency in the development of important pieces of certified software such as the C compiler of the CompCert project [78]. The extracted CompCert compiler is reliable and efficient, running only 15% slower than GCC 4 at optimisation level 2 (`gcc -O2`), a level of optimisation that was considered before to be highly unreliable.

Proof assistants can also be used to *formalise mathematical theories*: they not only provide a means of representing mathematical theories in a form amenable to computer processing, but their internal logic provides a language for reasoning about such theories. In the last decade, proof assistants have been used to verify extremely large and complicated proofs of recent mathematical results, sometimes requiring either intensive computations [60], [64] or intricate combinations of a multitude of mathematical theories [59]. But formalised mathematics is more than just proof checking and proof assistants can help with the organisation mathematical knowledge or even with the discovery of new constructions and proofs.

Unfortunately, the rigidity of the theory behind proof assistants impedes their expressiveness both as programming languages and as logical systems. For instance, a program extracted from Coq only uses a purely functional subset of OCaml, leaving behind important means of expression such as side-effects and objects. Limitations also appears in the formalisation of advanced mathematics: proof assistants do not cope well with classical axioms such as excluded middle and choice which are sometimes used crucially. The fact of the matter is that the development of proof assistants cannot be dissociated from a reflection on the nature of programs and proofs coming from the Curry-Howard correspondence. In the EPC Gallinette, we propose to address several drawbacks of proof assistants by pushing the boundaries of this correspondence.

In the 1970's, the Curry-Howard correspondence was seen as a perfect match between functional programs, intuitionistic logic, and Cartesian closed categories. It received several generalisations over the decades, and now it is more widely understood as a fertile correspondence between computation, logic, and algebra. Nowadays, the view of the Curry-Howard correspondence has evolved from a perfect match to a collection of theories meant to explain similar structures at work in logic and computation, underpinned by mathematical abstractions. By relaxing the requirement of a perfect match between programs and proofs, and instead emphasising the common foundations of both, the insights of the Curry-Howard correspondence may be extended to domains for which the requirements of programming and mathematics may in fact be quite different.

Consider the following two major theories of the past decades, which were until recently thought to be irreconcilable:

- **(Martin-Löf) Type theory:** introduced by Martin-Löf in 1971, this formalism [85] is both a programming language and a logical system. The central ingredient is the use of *dependent types* to allow fine-grained invariants to be expressed in program types. In 1985, Coquand and Huet developed a similar system called the *calculus of constructions*, which served as logical foundation of the first implementation of Coq. This kind of systems is still under active development, especially with the recent advent of homotopy type theory (HoTT) [107] which gives a new point of view on types and the notion of equality in type theory.

- **The theory of effects:** starting in the 1980's, Moggi [90] and Girard [57] put forward monads and co-monads as describing various compositional notions of computation. In this theory, programs can have side-effects (state, exceptions, input-output), logics can be non-intuitionistic (linear, classical), and different computational universes can interact (modal logics). Recently, the safe and automatic management of resources has also seen a coming of age (Rust, Modern C++) confirming the importance of linear logic for various programming concepts. It is now understood that the characteristic feature of the theory of effects is sensitivity to *evaluation order*, in contrast with type theory which is built around the assumption that evaluation order is irrelevant.

We now outline a series of scientific challenges aimed at understanding of type theory, effects, and their combination.

More precisely, three key axes of improvement have been identified:

1. Making the notion of equality closer to what is usually assumed when doing proofs on black board, with a balance between irrelevant equality for simple structures and equality up-to equivalences for more complex ones (Section 3.2). Such a notion of equality should allow one to implement traditional model transformations that enhance the logical power of the proof assistant using distinct compilation phases.
2. Advancing the foundations of effects within the Curry-Howard approach. The objective is to pave the way for the integration of effects in proof assistants and to prototype the corresponding implementation. This integration should allow for not only certified programming with effects, but also the expression of more powerful logics (Section 3.3).
3. Making more programming features (notably, object polymorphism) available in proof assistants, in order to scale to practical-sized developments. The objective is to enable programming styles closer to common practices. One of the key challenges here is to leverage gradual typing to dependent programming (Section 3.4).

To validate the new paradigms, we propose in Section 3.5 three particular application fields in which members of the team already have a strong expertise: code refactoring, constraint programming and symbolic computation.

3.2. Enhance the computational and logical power of proof assistants

The democratisation of proof assistants based on type theory has likely been impeded one central problem: the mismatch between the conception of equality in mathematics and its formalisation in type theory. Indeed, some basic principles that are used implicitly in mathematics—such as Church's principle of propositional extensionality, which says that two propositions are equal when they are logically equivalent—are not derivable in type theory. Even more problematically, from a computer science point of view, the basic concept of two functions being equal when they are equal at every "point" of their domain is also not derivable: rather, it must be added as an additional axiom. Of course, these principles are consistent with type theory so that working under the corresponding additional assumptions is safe. But the use of these assumptions in a definition potentially clutters its computational behaviour: since axioms are computational black boxes, computation gets stuck at the points of the code where they have been used.

We propose to investigate how expressive logical transformations such as forcing [70] and sheaf construction might be used to enhance the computational and logical power of proof assistants—with a particular emphasis on their implementation in the Coq proof assistant by the means of effective translations (or compilation phases). One of the main topics of this task, in connection to the ERC project CoqHoTT, is the integration in Coq of new concepts inspired by homotopy type theory [107] such as the univalence principle, and higher inductive types.

3.2.1. A definitional proof-irrelevant version of Coq.

In the Coq proof assistant, the sort **Prop** stands for the universe of types which are propositions. That is, when a term P has type **Prop**, the only relevant fact is whether P is inhabited (that is true) or not (that is false). This property, known as *proof irrelevance,* can be expressed formally as: $\forall x\ y : P, x = y$. Originally, the *raison d'être* of the sort **Prop** was to characterise types with no computational meaning with the intention that terms of such types could be erased upon extraction. However, the assumption that every element of **Prop** should be proof irrelevant has never been integrated to the system. Indeed, in Coq, proof irrelevance for the sort **Prop** is not incorporated into the theory: it is only compatible with it, in the sense that its assumption does not give rise to an inconsistent theory. In fact, the exact status of the sort **Prop** in Coq has never been entirely clarified, which explains in part this lack of integration. Homotopy type theory brings fresh thinking on this issue and suggests turning **Prop** into the collection of terms that a certain static inference procedure tags as proof irrelevant. The goal of this task is to integrate this insight in the Coq system and to implement a definitional proof-irrelevant version of the sort **Prop**.

3.2.2. Extend the Coq proof assistant with a computational version of univalence

The univalence principle is becoming widely accepted as a very promising avenue to provide new foundations for mathematics and type theory. However, this principle has not yet been incorporated into a proof assistant. Indeed, the very mathematical structures (known as ∞-groupoids) motivating the theory remain to this day an active area of research. Moreover, a correct and decidable type checking procedure for the whole theory raises both computational complexity and logical coherence issues. Observational type theory [32], as implemented in Epigram, provides a first-stage approximation to homotopy type theory, but only deals with functional extensionality and does not capture univalence. Coquand and his collaborators have obtained significant results on the computational meaning of univalence using cubical sets [39], [45]. Bickford has initiated a promising formalisation work [0] in the NuPRL system. However, a complete formalisation in intensional type theory remains an open problem.

Hence a major objective is to achieve a complete internalisation of univalence in intensional type theory, including an integration to a new version of Coq. We will strive to keep compatibility with previous versions, in particular from a performance point of view. Indeed, the additional complexity of homotopy type theory should not induce an overhead in the type checking procedure used by the software if we want our new framework to become rapidly adopted by the community. Concretely, we will make sure that the compilation time of Coq's Standard Library will be of the same order of magnitude.

3.2.3. Extend the logical power of type theory without axioms in a modular way

Extending the power of a logic using model transformations (*e.g.,* forcing transformation [71], [70] or the sheaf construction [100]) is a classic topic of mathematical logic [46], [76]. However, these ideas have not been much investigated in the setting of type theory, even though they may provide a useful framework for extending the logical power of proof assistant in a modular way. There is a good reason for this: with a syntactic notion of equality, the underlying structure of type theory does not conform to the structure of topos used in mathematical logic. A direct incorporation of the standard techniques is therefore not possible. However, a univalent notion of equality brings type theory closer to the required algebraic structure, as it corresponds to the notion of ∞-topos recently studied by Lurie [83]. The goal of this task is to revisit model transformations in the light of the univalence principle, and to obtain in this way new internal transformations in type theory which can in turn be seen as compilation phases. The general notion of an internal syntactical translation has already been investigated in the team [40].

3.2.4. Methodology: Extending type theory with different compilation phases

The Gallinette project advocates the use of distinct compilation phases as a methodology for the design of a new generation of proof assistants featuring modular extensions of a core logic. The essence of a compiler is the separation of the complexity of a translation process into modular stages, and the organization of their

[0] http://www.nuprl.org/wip/Mathematics/cubical!type!theory/index.html

Figure 1. Multiple compilation phases to increase the logical and computational power of Coq.

re-composition. This idea finds a natural application in the design of complex proof assistants (Figure 1). For instance, the definition of type classes in Coq follows this pattern, and is morally given by the means of a translation into a type-class free kernel. More recently, a similar approach by compilation stages, using the forcing transformation, was used to relax the strict positivity condition guarding inductive types [71], [70]. We believe that this flavour of compilation-based strategies offers a promising direction of investigation for the propose of defining a decidable type checking algorithm for HoTT.

3.3. Semantic and logical foundations for effects in proof assistants based on type theory

We propose the incorporation of effects in the theory of proof assistants at a foundational level. Not only would this allow for certified programming with effects, but it would moreover have implications for both semantics and logic.

We mean *effects* in a broad sense that encompasses both Moggi's monads [90] and Girard's linear logic [57]. These two seminal works have given rise to respective theories of effects (monads) and resources (co-monads). Recent advances, however have unified these two lines of thought: it is now clear that the defining feature of effects, in the broad sense, is sensitivity to evaluation order [79], [50].

In contrast, the type theory that forms the foundations of proof assistants is based on pure λ calculus and is built on the assumption that evaluation order is irrelevant. Evaluation order is therefore the blind spot of type theory. In Moggi [91], integrating the dependent types of type theory with monads is *"the next difficult step [...] currently under investigation"*.

Any realistic program contains effects: state, exceptions, input-output. More generally, evaluation order may simply be important for complexity reasons. With this in mind, many works have focused on certified programming with effects: notably Ynot [95], and more recently $F^{\star}$ [105] and Idris [41], which propose various ways for encapsulating effects and restricting the dependency of types on effectful terms. Effects are either specialised, such as the monads with Hoare-style pre- and post-conditions found in Ynot or $F^{\star}$, or more general, such as the algebraic effects implemented in Idris. But whereas there are several experiments and projects pursuing the certification of programs with effects, each making its own choices on how effects and dependency should be merged, there is on the other hand a deficit of logical and semantic investigations.

We propose to develop the foundations of a type theory with effects taking into account the logical and semantic aspects, and to study their practical and theoretical consequences. A type theory that integrates effects would have logical, algebraic and computational implications when viewed through the Curry-Howard correspondence. For instance, effects such as control operators establish a link with classical proof theory [62]. Indeed, control operators provide computational interpretations of type isomorphisms such as $A \cong \neg\neg A$ and $\neg\forall x A \cong \exists x \neg A$ (e.g. [92]), whereas the conventional wisdom of type theory holds that such axioms are non-constructive (this is for instance the point of view that has been advocated so far in homotopy type theory [107]). Another example of an effect with logical content is state (more precisely memoization) which is used to provide constructive content to the classical dependent axiom of choice [38], [74], [66]. In the long term, a whole body of literature on the constructive content of classical proofs is to be explored and integrated, providing rich sources of inspiration: Kohlenbach's proof mining [73] and Simpson's reverse mathematics [103], for instance, are certainly interesting to investigate from the Curry-Howard perspective.

The goal is to develop a type theory with effects that accounts both for practical experiments in certified programming, and for clues from denotational semantics and logical phenomena, in a unified setting.

3.3.1. Models for integrating effects with dependent types

A crucial step is the integration of dependent types with effects, a topic which has remained *"currently under investigation"* [91] ever since the beginning. The difficulty resides in expressing the dependency of types on terms that can perform side-effects during the computation. On the side of denotational semantics, several extensions of categorical models for effects with dependent types have been proposed [29], [108] using axioms that should correspond to restrictions in terms of expressivity but whose practical implications, however, are not immediately transparent. On the side of logical approaches [66], [67], [77], [89], one first considers a drastic restriction to terms that do not compute, which is then relaxed by semantic means. On the side of systems for certified programming such as $\mathrm{F}^{\star}$, the type system ensures that types only depend on pure and terminating terms.

Thus, the recurring idea is to introduce restrictions on the dependency in order to establish an encapsulation of effects. In our approach, we seek a principled description of this idea by developing the concept of *semantic value* (thunkables, linears) which arose from foundational considerations [56], [102], [93] and whose relevance was highlighted in recent works [80], [99]. The novel aspect of our approach is to seek a proper extension of type theory which would provide foundations for a classical type theory with axiom of choice in the style of Herbelin [66], but which moreover could be generalised to effects other than just control by exploiting an abstract and adaptable notion of semantic value.

3.3.2. Intuitionistic depolarisation

In our view, the common idea that evaluation order does not matter for pure and termination computations should serve as a bridge between our proposals for dependent types in the presence of effects and traditional type theory. Building on the previous goal, we aim to study the relationship between semantic values, purity, and parametricity theorems [101], [58]. Our goal is to characterise parametricity as a form of intuitionistic *depolarisation* following the method by which the first game model of full linear logic was given (Melliès [86], [87]). We have two expected outcomes in mind: enriching type theory with intensional content without losing its properties, and giving an explanation of the dependent types in the style of Idris and $\mathrm{F}^{\star}$ where purity- and termination-checking play a role.

3.3.3. Developing the rewriting theory of calculi with effects

An integrated type theory with effects requires an understanding of evaluation order from the point of view of rewriting. For instance, rewriting properties can entail the decidability of some conversions, allowing the automation of equational reasoning in types [27]. They can also provide proofs of computational consistency (that terms are not all equivalent) by showing that extending calculi with new constructs is conservative [104]. In our approach, the λ-calculus is replaced by a calculus modelling the evaluation in an abstract machine [51]. We have shown how this approach generalises the previous semantic and proof-theoretic approaches [33], [79], [81], and overcomes their shortcomings [94].

One goal is to prove computational consistency or decidability of conversions purely using advanced rewriting techniques following a technique introduced in [104]. Another goal is the characterisation of weak reductions: extensions of the operational semantics to terms with free variables that preserve termination, whose iteration is equivalent to strong reduction [28], [54]. We aim to show that such properties derive from generic theorems of higher-order rewriting [110], so that weak reduction can easily be generalised to richer systems with effects.

3.3.4. Direct models and categorical coherence

Proof theory and rewriting are a source of *coherence theorems* in category theory, which show how calculations in a category can be simplified with an embedding into a structure with stronger properties [84], [75]. We aim to explore such results for categorical models of effects [79], [50]. Our key insight is to consider the reflection between *indirect and direct models* [56], [93] as a coherence theorem: it allows us to embed the traditional models of effects into structures for which the rewriting and proof-theoretic techniques from the previous section are effective.

Building on this, we are further interested in connecting operational semantics to 2-category theory, in which a second dimension is traditionally considered for modelling conversions of programs rather than equivalences. This idea has been successfully applied for the λ-calculus [72], [68] but does not scale yet to more realistic models of computation. In our approach, it has already been noticed that the expected symmetries coming from categorical dualities are better represented, motivating a new investigation into this long-standing question.

3.3.5. Models of effects and resources

The unified theory of effects and resources [50] prompts an investigation into the semantics of safe and automatic resource management, in the style of Modern C++ and Rust. Our goal is to show how advanced semantics of effects, resources, and their combination arise by assembling elementary blocks, pursuing the methodology applied by Melliès and Tabareau in the context of continuations [88]. For instance, by combining control flow (exceptions, return) with linearity allows us to describe in a precise way the "Resource Acquisition Is Initialisation" idiom in which the resource safety is ensured with scope-based destructors. A further step would be to reconstruct uniqueness types and borrowing using similar ideas.

3.4. Language extensions for the scaling of proof assistants

The development of tools to construct software systems that respect a given specification is a major challenge of current and future research in computer science. Certified programming with dependent types has recently attracted a lot of interest, and Coq is the *de facto* standard for such endeavours, with an increasing number of users, pedagogical resources, and large-scale projects. Nevertheless, significant work remains to be done to make Coq more usable from a software engineering point of view. The Gallinette team proposes to make progress on three lines of work: (i) the development of gradual certified programming, (ii) the integration of imperative features and object polymorphism in Coq, and (iii) the development of robust tactics for proof engineering for the scaling of formalised libraries.

3.4.1. Gradual Certified Programming

One of the main issues faced by a programmer starting to internalise in a proof assistant code written in a more permissive world is that type theory is constrained by a strict type discipline which lacks flexibility. Concretely, as soon that you start giving more a precise type/specification to a function, the rest of the code interacting with this functions needs to be more precise too. To address this issue, the Gallinette team will put strong efforts into the development of gradual typing in type theory to allow progressive integration of code that comes from a more permissive world.

Indeed, on the way to full verification, programmers can take advantage of a gradual approach in which some properties are simply asserted instead of proven, subject to dynamic verification. Tabareau and Tanter have made preliminary progress in this direction [106]. This work, however, suffers from a number of limitations, the most important being the lack of a mechanism for handling the possibility of runtime errors within Coq. Instead of relying on axioms, this project will explore the application of Section 3.3 to embed effects in Coq.

This way, instead of postulating axioms for parts of the development that are too hard/marginal to be dealt with, the system adds dynamic checks. Then, after extraction, we get a program that corresponds to the initial program but with dynamic check for parts that have not been proven, ensuring that the program will raise an error instead of going outside its specification.

This will yield new foundations of gradual certified programming, both more expressive and practical. We will also study how to integrate previous techniques with the extraction mechanism of Coq programs to OCaml, in order to exploit the exception mechanism of OCaml.

3.4.2. Imperative features and object polymorphism in the Coq proof assistant

3.4.2.1. Imperative features.

Abstract data types (ADTs) become useful as the size of programs grows since they provide for a modular approach, allowing abstractions about data to be expressed and then instantiated. Moreover, ADTs are natural concepts in the calculus of inductive constructions. But while it is easy to declare an ADT, it is often difficult to implement an efficient one. Compare this situation with, for example, Okasaki's purely functional data structures [96] which implement ADTs like queues in languages with imperative features. Of course, Okasaki's queues enforce some additional properties for free, such as persistence, but the programmer may prefer to use and to study a simpler implementation without those additional properties. Also in certified symbolic computation (see 3.5.3), an efficient functional implementation of ADTs is often not available, and efficiency is a major challenge in this area. Relying on the theoretical work done in 3.3 , we will equip Coq with imperative features and we will demonstrate how they can be used to provide efficient implementations of ADTs. However, it is also often the case that imperative implementation are hard-to-reason-on, requiring for instance the use of separation logic. But in that case, we could take benefice of recent works on integration of separation logic in the Coq proof assistant and in particular the Iris project http://iris-project.org/.

3.4.2.2. Object polymorphism.

Object-oriented programming has evolved since its foundation based on the representation of computations as an exchange of messages between objects. In modern programming languages like Scala, which aims at a synthesis between object-oriented and functional programming, object-orientation concretely results in the use of hierarchies of interfaces ordered by the subtyping relation and the definition of interface implementations that can interoperate. As observed by Cook and Aldrich [48], [31], interoperability can be considered as the essential feature of objects and is a requirement for many modern frameworks and ecosystems: it means that two different implementations of the same interface can interoperate.

Our objective is to provide a representation of object-oriented programs, by focusing on subtyping and interoperability.

For subtyping, the natural solution in type theory is coercive subtyping [82], as implemented in Coq, with an explicit operator for coercions. This should lead to a shallow embedding, but has limitations: indeed, while it allows subtyping to be faithfully represented, it does not provide a direct means to represent union and intersection types, which are often associated with subtyping (for instance intersection types are present in Scala). A more ambitious solution would be to resort to subsumptive subtyping (or semantic subtyping [55]): in its more general form, a type algebra is extended with boolean operations (union, intersection, complementing) to get a boolean algebra with operators (the original type constructors). Subtying is then interpreted as the natural partial order of the boolean algebra.

We propose to use the type class machinery of Coq to implement semantic subtyping for dependent type theory. Using type class resolution, we can emulate inference rules of subsumptive subtyping without modifying Coq internally. This has also another advantage. As subsumptive subtyping for dependent types should be undecidable in general, using type class resolution allows for an incomplete yet extensible decision procedure.

3.4.3. Robust tactics for proof engineering for the scaling of formalised libraries

When developing certified software, a major part of the effort is spent not only on writing proof scripts, but on *rewriting* them, either for the purpose of code maintenance or because of more significant changes in the base

definitions. Regrettably, proof scripts suffer more often than not from a bad programming style, and too many proof developers casually neglect the most elementary principles of well-behaved programmers. As a result, many proof scripts are very brittle, user-defined tactics are often difficult to extend, and sometimes even lack a clear specification. Formal libraries are thus generally very fragile pieces of software. One reason for this unfortunate situation is that proof engineering is very badly served by the tools currently available to the users of the Coq proof assistant, starting with its tactic language. One objective of the Gallinette team is to develop better tools to write proof scripts.

Completing and maintaining a large corpus of formalised mathematics requires a well-designed tactic language. This language should both accommodate the possible specific needs of the theories at stake, and help with diagnostics at refactoring time. Coq's tactic language is in fact two-leveled. First, it includes a basic tactic language, to organise the deductive steps in a proof script and to perform the elementary bureaucracy. Its second layer is a meta-programming language, which allows user to defined their own new tactics at toplevel. Our first direction of work consists in the investigation of the appropriate features of the *basic tactic language*. For instance, the design of the Ssreflect tactic language, and its support for the small scale reflection methodology [61], has been a key ingredient in at least two large scale formalisation endeavours: the Four Colour Theorem [60] and of the Odd Order Theorem [59]. Building on our experience with the Ssreflect tactic language, we will contribute to the ongoing work on the basic tactic language for Coq. The second objective of this task is to contribute to the design of a *typed tactic language*. In particular, we will build on the work of Ziliani and his collaborators [109], extending it with reasoning about the effects that tactics have on the "state of a proof" (e.g. number of sub-goals, metavariables in context). We will also develop a novel approach for incremental type checking of proof scripts, so that programmers gain access to a richer discovery- engineering interaction with the proof assistant.

3.5. Practical experiments

The first three axes of the EPC Gallinette aim at developing a new generation of proof assistants. But we strongly believe that foundational investigations must go hand in hand with practical experiments. Therefore, we expect to benefit from existing expertise and collaborations in the team to experiment our extensions of Coq on real world developments. It should be noticed that those practical experiments are strongly guided by the deep history of research on software engineering of team members.

3.5.1. Certified Code Refactoring

In the context of refactoring of C programs, we intend to formalise program transformations that are written in an imperative style to test the usability of our addition of effects in the proof assistant. This subject has been chosen based on the competence of members of the team.

We are currently working on the formalisation of refactoring tools in Coq [44]. Automatic refactoring of programs in industrial languages is difficult because of the large number of potential interactions between language features that are difficult to predict and to test. Indeed, all available refactoring tools suffer from bugs : they fail to ensure that the generated program has the same behaviour as the input program. To cope with that difficulty, we have chosen to build a refactoring tool with Coq : a program transformation is written in the Coq programming language, then proven correct on all possible inputs, and then an OCaml executable program is generated by the platform. We rely on the CompCert C formalisation of the C language. CompCert is currently the most complete formalisation of an industrial language, which justifies that choice. We have three goals in that project :

- Build a refactoring tool that programmers can rely on and make it available in a popular platform (such as Eclipse, IntelliJ or Frama-C).
- Explore large, drastic program transformations such as replacing a design architecture for an other one, by applying a sequence of small refactoring operations (as we have done for Java and Haskell programs before [47], [43], [30]), while ensuring behaviour preservation.
- Explore the use of enhancements of proof systems on large developments. For instance, refactoring tools are usually developed in the imperative/object paradigm, so the extension of Coq with side effects or with object features proposed in the team can find a direct use-case here.

3.5.2. Certified Constraint Programming

We plan to make use of the internalisation of the object-oriented paradigm in the context of constraint programming. Indeed, this domain is made of very complex algorithms that are often developed using object-oriented programming (as it is the case for instance for CHOCO, which is developed in the Tasc Group at IMT Atlantique, Nantes). We will in particular focus on filtering algorithms in constraint solvers, for which research publications currently propose new algorithms with manual proofs. Their formalisation in Coq is challenging. Another interesting part of constraint solving to formalise is the part that deals with program generation (as opposed to extraction). However, when there are numerous generated pieces of code, it is not realistic to prove their correctness manually, and it can be too difficult to prove the correctness of a generator. So we intend to explore a middle path that consists in generating a piece of code along with its corresponding proof (script or proof term). A target application could be interval constraints (for instance Allen interval algebra or region connection calculus) that can generate thousands of specialised filtering algorithms for a small number of variables [36].

Finally, Rémi Douence has already worked (articles publishing [63], [97], [53], PhD Thesis advising [98]) with different members of the Tasc team. Currently, he supervises with Nicolas Beldiceanu the PhD Thesis of Ekaterina Arafailova in the Tasc team. She studies finite transducers to model time-series constraints [37], [35], [34]. This work requires proofs, manually done for now, we would like to explore when these proofs could be mechanised.

3.5.3. Certified Symbolic Computation

We will investigate how the addition of effects in the Coq proof assistant can facilitate the marriage of computer algebra with formal proofs. Computer algebra systems on one hand, and proof assistants on the other hand, are both designed for doing mathematics with the help of a computer, by the means of symbolic computations. These two families of systems are however very different in nature: computer algebra systems allow for implementations faithful to the theoretical complexity of the algorithms, whereas proof assistants have the expressiveness to specify exactly the semantic of the data-structures and computations.

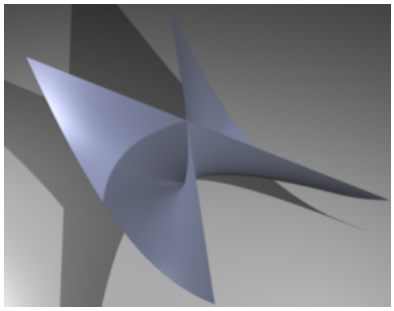
Experiments have been run that link computer algebra systems with Coq [52], [42]. These bridges rely on the implementation of formal proof-producing core algorithms like normalisation procedures. Incidentally, they require non trivial maintenance work to survive the evolution of both systems. Other proof assistants like the Isabelle/HOL system make use of so-called reflection schemes: the proof assistant can produce code in an external programming language like SML, but also allows to import the values output by these extracted programs back inside the formal proofs. This feature extends the trusted base of code quite significantly but it has been used for major achievements like a certified symbolic/numeric ODE solver [69].

We would like to bring Coq closer to the efficiency and user-friendliness of computer algebra systems: for now it is difficult to use the Coq programming language so that certified implementations of computer algebra algorithms have the right, observable, complexity when they are executed inside Coq. We see the addition of effects to the proof assistant as an opportunity to ease these implementations, for instance by making use of caching mechanisms or of profiling facilities. Such enhancements should enable the verification of computation-intensive mathematical proofs that are currently beyond reach, like the validation of Helfgott's proof of the weak Goldbach conjecture [65].

GAMBLE Project-Team

3. Research Program

3.1. Non-linear computational geometry



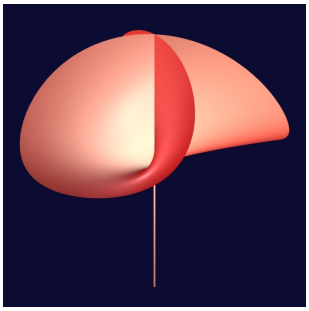


Figure 1. Two views of the Whitney umbrella (on the left, the "stick" of the umbrella, i.e., the negative z-axis, is missing). Right picture from [Wikipedia], left picture from [Lachaud et al.].

As mentioned above, curved objects are ubiquitous in real world problems and in computer science and, despite this fact, there are very few problems on curved objects that admit robust and efficient algorithmic solutions without first discretizing the curved objects into meshes. Meshing curved objects induces a loss of accuracy which is sometimes not an issue but which can also be most problematic depending on the application. In addition, discretization induces a combinatorial explosion which could cause a loss in efficiency compared to a direct solution on the curved objects (as our work on quadrics has demonstrated with flying colors [50], [51], [52], [54], [58]). But it is also crucial to know that even the process of computing meshes that approximate curved objects is far from being resolved. As a matter of fact there is no algorithm capable of computing in practice meshes with certified topology of even rather simple singular 3D surfaces, due to the high constants in the theoretical complexity and the difficulty of handling degenerate cases. Part of the difficulty comes from the unintuitive fact that the structure of an algebraic object can be quite complicated, as depicted in the Whitney umbrella (see Figure 1), surface of equation $x^2 = y^2 z$ on which the origin (the "special" point of the surface) is a vertex of the arrangement induced by the surface while the singular locus is simply the whole z-axis. Even in 2D, meshing an algebraic curve with the correct topology, that is in other words producing a correct drawing of the curve (without knowing where the domain of interest is), is a very difficult problem on which we have recently made important contributions [37], [38], [59].

It is thus to be understood that producing practical robust and efficient algorithmic solutions to geometric problems on curved objects is a challenge on all and even the most basic problems. The basicness and fundamentality of two problems we mentioned above on the intersection of 3D quadrics and on the drawing in a topologically certified way of plane algebraic curves show rather well that the domain is still in its infancy. And it should be stressed that these two sets of results were not anecdotal but flagship results produced during the lifetime of the VEGAS team (the team preceding GAMBLE).

There are many problems in this theme that are expected to have high long-term impacts. Intersecting NURBS (Non-uniform rational basis splines) in a certified way is an important problem in computer-aided design and manufacturing. As hinted above, meshing objects in a certified way is important when topology matters. The 2D case, that is essentially drawing plane curves with the correct topology, is a fundamental problem with

far-reaching applications in research or R&D. Notice that on such elementary problems it is often difficult to predict the reach of the applications; as an example, we were astonished by the scope of the applications of our software on 3D quadric intersection [0] which was used by researchers in, for instance, photochemistry, computer vision, statistics and mathematics.

3.2. Non-Euclidean computational geometry

Figure 2. Left: 3D mesh of a gyroid (triply periodic surface) [61]. Right: Simulation of a periodic Delaunay triangulation of the hyperbolic plane [33].

Triangulations, in particular Delaunay triangulations, in the *Euclidean space* $\mathbb{R}^d$ have been extensively studied throughout the 20th century and they are still a very active research topic. Their mathematical properties are now well understood, many algorithms to construct them have been proposed and analyzed (see the book of Aurenhammer *et al.* [32]). Some members of GAMBLE have been contributing to these algorithmic advances (see, e.g. [36], [68], [47], [35]); they have also contributed robust and efficient triangulation packages through the state-of-the-art Computational Geometry Algorithms Library CGAL whose impact extends far beyond computational geometry. Application fields include particle physics, fluid dynamics, shape matching, image processing, geometry processing, computer graphics, computer vision, shape reconstruction, mesh generation, virtual worlds, geophysics, and medical imaging. [0]

It is fair to say that little has been done on non-Euclidean spaces, in spite of the large number of questions raised by application domains. Needs for simulations or modeling in a variety of domains [0] ranging from the infinitely small (nuclear matter, nano-structures, biological data) to the infinitely large (astrophysics) have led us to consider 3D periodic Delaunay triangulations, which can be seen as Delaunay triangulations in the 3D *flat torus*, quotient of $\mathbb{R}^3$ under the action of some group of translations [42]. This work has already yielded a fruitful collaboration with astrophysicists [55], [69] and new collaborations with physicists are emerging. To the best of our knowledge, our CGAL package [41] is the only publicly available software that computes Delaunay triangulations of a 3D flat torus, in the special case where the domain is cubic. This case, although restrictive, is already useful. [0] We have also generalized this algorithm to the case of general d-dimensional compact flat manifolds [43]. As far as non-compact manifolds are concerned, past approaches, limited to the two-dimensional case, have stayed theoretical [60].

Interestingly, even for the simple case of triangulations on the *sphere*, the software packages that are currently available are far from offering satisfactory solutions in terms of robustness and efficiency [40].

[0] QI: web.
[0] See Projects using CGAL for details.
[0] See CGAL Prospective Workshop on Geometric Computing in Periodic Spaces, Subdivide and Tile: Triangulating spaces for understanding the world, Computational geometry in non-Euclidean spaces, Shape Up 2015 : Exercises in Materials Geometry and Topology
[0] See examples at Projects using CGAL

Moreover, while our solution for computing triangulations in hyperbolic spaces can be considered as ultimate [33], the case of *hyperbolic manifolds* has hardly been explored. Hyperbolic manifolds are quotients of a hyperbolic space by some group of hyperbolic isometries. Their triangulations can be seen as hyperbolic periodic triangulations. Periodic hyperbolic triangulations and meshes appear for instance in geometric modeling [62], neuromathematics [45], or physics [65]. Even the case of the Bolza surface (a surface of genus 2, whose fundamental domain is the regular octagon in the hyperbolic plane) shows mathematical difficulties [34], [57].

3.3. Probability in computational geometry

In most computational geometry papers, algorithms are analyzed in the worst-case setting. This often yields too pessimistic complexities that arise only in pathological situations that are unlikely to occur in practice. On the other hand, probabilistic geometry provides analyses with great precision [63], [64], [39], but using hypotheses with much more randomness than in most realistic situations. We are developing new algorithmic designs improving state-of-the-art performance in random settings that are not overly simplified and that can thus reflect many realistic situations.

Twelve years ago, smooth analysis was introduced by Spielman and Teng analyzing the simplex algorithm by averaging on some noise on the data [67] (and they won the Gödel prize). In essence, this analysis smoothes the complexity around worst-case situations, thus avoiding pathological scenarios but without considering unrealistic randomness. In that sense, this method makes a bridge between full randomness and worst case situations by tuning the noise intensity. The analysis of computational geometry algorithms within this framework is still embryonic. To illustrate the difficulty of the problem, we started working in 2009 on the smooth analysis of the size of the convex hull of a point set, arguably the simplest computational geometry data structure; then, only one very rough result from 2004 existed [46] and we only obtained in 2015 breakthrough results, but still not definitive [49], [48], [53].

Another example of a problem of different flavor concerns Delaunay triangulations, which are rather ubiquitous in computational geometry. When Delaunay triangulations are computed for reconstructing meshes from point clouds coming from 3D scanners, the worst-case scenario is, again, too pessimistic and the full randomness hypothesis is clearly not adapted. Some results exist for "good samplings of generic surfaces" [31] but the big result that everybody wishes for is an analysis for random samples (without the extra assumptions hidden in the "good" sampling) of possibly non-generic surfaces.

Trade-offs between full randomness and worst case may also appear in other forms such as dependent distributions, or random distributions conditioned to be in some special configurations. Simulating these kinds of geometric distributions is currently out of reach for more than a few hundred points [56] although it has practical applications in physics or networks.

3.4. Discrete geometric structures

Our work on discrete geometric structures develops in several directions, each one probing a different type of structure. Although these objects appear unrelated at first sight, they can be tackled by the same set of probabilistic and topological tools.

A first research topic is the study of *Order types*. Order types are combinatorial encodings of finite (planar) point sets, recording for each triple of points the orientation (clockwise or counterclockwise) of the triangle they form. This already determines properties such as convex hulls or half-space depths, and the behaviour of algorithms based on orientation predicates. These properties for all (infinitely many) n-point sets can be studied through the finitely many order types of size n. Yet, this finite space is poorly understood: its estimated size leaves an exponential margin of error, no method is known to sample it without concentrating on a vanishingly small corner, the effect of pattern exclusion or VC dimension-type restrictions are unknown. These are all directions we actively investigate.

A second research topic is the study of *Embedded graphs and simplicial complexes.* Many topological structures can be effectively discretized, for instance combinatorial maps record homotopy classes of embedded graphs and simplicial complexes represent a large class of topological spaces. This raises many structural and algorithmic questions on these discrete structures; for example, given a closed walk in an embedded graph, can we find a cycle of the graph homotopic to that walk? (The complexity status of that problem is unknown.) Going in the other direction, some purely discrete structures can be given an associated topological space that reveals some of their properties (*e.g.* the Nerve theorem for intersection patterns). An open problem is for instance to obtain fractional Helly theorems for set system of bounded topological complexity.

Another research topic is that of *Sparse inclusion-exclusion formulas.* For any family of sets $A_1, A_2, ..., A_n$, by the principle of inclusion-exclusion we have

$$\mathbb{1}_{\bigcup_{i=1}^{n} A_i} = \sum_{I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} \mathbb{1}_{\bigcap_{i \in I} A_i} \tag{1}$$

where $\mathbb{1}_X$ is the indicator function of X. This formula is universal (it applies to any family of sets) but its number of summands grows exponentially with the number n of sets. When the sets are balls, the formula remains true if the summation is restricted to the regular triangulation; we proved that similar simplifications are possible whenever the Venn diagram of the A_i is sparse. There is much room for improvements, both for general set systems and for specific geometric settings. Another interesting problem (the subject of the PhD thesis of Galatée Hemery) is to combine these simplifications with the inclusion-exclusion algorithms developed, for instance, for graph coloring.

GRACE Project-Team

3. Research Program

3.1. Algorithmic Number Theory

Participants: Luca de Feo, François Morain, Benjamin Smith, Mathilde de La Morinerie, Antonin Leroux, Guénaël Renault.

Algorithmic Number Theory is concerned with replacing special cases with general algorithms to solve problems in number theory. In the Grace project, it appears in three main threads:

- fundamental algorithms for integers and polynomials (including primality and factorization);
- algorithms for finite fields (including discrete logarithms);
- algorithms for algebraic curves.

Clearly, we use computer algebra in many ways. Research in cryptology has motivated a renewed interest in Algorithmic Number Theory in recent decades—but the fundamental problems still exist *per se*. Indeed, while algorithmic number theory application in cryptanalysis is epitomized by applying factorization to breaking RSA public key, many other problems, are relevant to various area of computer science. Roughly speaking, the problems of the cryptological world are of bounded size, whereas Algorithmic Number Theory is also concerned with asymptotic results.

3.2. Arithmetic Geometry: Curves and their Jacobians

Participants: Luca de Feo, François Morain, Benjamin Smith, Mathilde de La Morinerie, Antonin Leroux.

Theme: Arithmetic Geometry: Curves and their Jacobians *Arithmetic Geometry* is the meeting point of algebraic geometry and number theory: that is, the study of geometric objects defined over arithmetic number systems (such as the integers and finite fields). The fundamental objects for our applications in both coding theory and cryptology are curves and their Jacobians over finite fields.

An algebraic *plane curve* $\mathcal{X}$ over a field $\mathbf{K}$ is defined by an equation

$$\mathcal{X} : F_{\mathcal{X}}(x, y) = 0 \quad \text{where } F_{\mathcal{X}} \in \mathbf{K}[x, y].$$

(Not every curve is planar—we may have more variables, and more defining equations—but from an algorithmic point of view, we can always reduce to the plane setting.) The *genus* $g_{\mathcal{X}}$ of $\mathcal{X}$ is a non-negative integer classifying the essential geometric complexity of $\mathcal{X}$; it depends on the degree of $F_{\mathcal{X}}$ and on the number of singularities of $\mathcal{X}$. The curve $\mathcal{X}$ is associated in a functorial way with an algebraic group $J_{\mathcal{X}}$, called the *Jacobian* of $\mathcal{X}$. The group $J_{\mathcal{X}}$ has a geometric structure: its elements correspond to points on a $g_{\mathcal{X}}$-dimensional projective algebraic group variety. Typically, we do not compute with the equations defining this projective variety: there are too many of them, in too many variables, for this to be convenient. Instead, we use fast algorithms based on the representation in terms of classes of formal sums of points on $\mathcal{X}$.

The simplest curves with nontrivial Jacobians are curves of genus 1, known as *elliptic curves*; they are typically defined by equations of the form $y^2 = x^3 + Ax + B$. Elliptic curves are particularly important given their central role in public-key cryptography over the past two decades. Curves of higher genus are important in both cryptography and coding theory.

3.3. Curve-Based cryptology

Participants: Luca de Feo, François Morain, Benjamin Smith, Mathilde de La Morinerie, Antonin Leroux.

Theme: Curve-Based Cryptology

Jacobians of curves are excellent candidates for cryptographic groups when constructing efficient instances of public-key cryptosystems. Diffie–Hellman key exchange is an instructive example.

Suppose Alice and Bob want to establish a secure communication channel. Essentially, this means establishing a common secret *key*, which they will then use for encryption and decryption. Some decades ago, they would have exchanged this key in person, or through some trusted intermediary; in the modern, networked world, this is typically impossible, and in any case completely unscalable. Alice and Bob may be anonymous parties who want to do e-business, for example, in which case they cannot securely meet, and they have no way to be sure of each other's identities. Diffie–Hellman key exchange solves this problem. First, Alice and Bob publicly agree on a cryptographic group G with a generator P (of order N); then Alice secretly chooses an integer a from $[1..N]$, and sends aP to Bob. In the meantime, Bob secretly chooses an integer b from $[1..N]$, and sends bP to Alice. Alice then computes $a(bP)$, while Bob computes $b(aP)$; both have now computed abP, which becomes their shared secret key. The security of this key depends on the difficulty of computing abP given P, aP, and bP; this is the Computational Diffie–Hellman Problem (CDHP). In practice, the CDHP corresponds to the Discrete Logarithm Problem (DLP), which is to determine a given P and aP.

This simple protocol has been in use, with only minor modifications, since the 1970s. The challenge is to create examples of groups G with a relatively compact representation and an efficiently computable group law, and such that the DLP in G is hard (ideally approaching the exponential difficulty of the DLP in an abstract group). The Pohlig–Hellman reduction shows that the DLP in G is essentially only as hard as the DLP in its largest prime-order subgroup. We therefore look for compact and efficient groups of prime order.

The classic example of a group suitable for the Diffie–Hellman protocol is the multiplicative group of a finite field $\mathbf{F}_q$. There are two problems that render its usage somewhat less than ideal. First, it has too much structure: we have a subexponential Index Calculus attack on the DLP in this group, so while it is very hard, the DLP falls a long way short of the exponential difficulty of the DLP in an abstract group. Second, there is only one such group for each q: its subgroup treillis depends only on the factorization of $q-1$, and requiring $q-1$ to have a large prime factor eliminates many convenient choices of q.

This is where Jacobians of algebraic curves come into their own. First, elliptic curves and Jacobians of genus 2 curves do not have a subexponential index calculus algorithm: in particular, from the point of view of the DLP, a generic elliptic curve is currently *as strong as* a generic group of the same size. Second, they provide some diversity: we have many degrees of freedom in choosing curves over a fixed $\mathbf{F}_q$, with a consequent diversity of possible cryptographic group orders. Furthermore, an attack which leaves one curve vulnerable may not necessarily apply to other curves. Third, viewing a Jacobian as a geometric object rather than a pure group allows us to take advantage of a number of special features of Jacobians. These features include efficiently computable pairings, geometric transformations for optimised group laws, and the availability of efficiently computable non-integer endomorphisms for accelerated encryption and decryption.

3.4. Algebraic Coding Theory

Participants: Daniel Augot, Alain Couvreur, Françoise Levy-Dit-Vehel, Maxime Roméas, Sarah Bordage, Adrien Hauteville, Isabella Panaccione.

Theme: Coding theory

Coding Theory studies originated with the idea of using redundancy in messages to protect against noise and errors. The last decade of the 20th century has seen the success of so-called iterative decoding methods, which enable us to get very close to the Shannon capacity. The capacity of a given channel is the best achievable transmission rate for reliable transmission. The consensus in the community is that this capacity is more easily reached with these iterative and probabilistic methods than with algebraic codes (such as Reed–Solomon codes).

However, algebraic coding is useful in settings other than the Shannon context. Indeed, the Shannon setting is a random case setting, and promises only a vanishing error probability. In contrast, the algebraic Hamming approach is a worst case approach: under combinatorial restrictions on the noise, the noise can be adversarial, with strictly zero errors.

These considerations are renewed by the topic of list decoding after the breakthrough of Guruswami and Sudan at the end of the nineties. List decoding relaxes the uniqueness requirement of decoding, allowing a small list of candidates to be returned instead of a single codeword. List decoding can reach a capacity close to the Shannon capacity, with zero failure, with small lists, in the adversarial case. The method of Guruswami and Sudan enabled list decoding of most of the main algebraic codes: Reed–Solomon codes and Algebraic–Geometry (AG) codes and new related constructions "capacity-achieving list decodable codes". These results open the way to applications again adversarial channels, which correspond to worst case settings in the classical computer science language.

Another avenue of our studies is AG codes over various geometric objects. Although Reed–Solomon codes are the best possible codes for a given alphabet, they are very limited in their length, which cannot exceed the size of the alphabet. AG codes circumvent this limitation, using the theory of algebraic curves over finite fields to construct long codes over a fixed alphabet. The striking result of Tsfasman–Vladut–Zink showed that codes better than random codes can be built this way, for medium to large alphabets. Disregarding the asymptotic aspects and considering only finite length, AG codes can be used either for longer codes with the same alphabet, or for codes with the same length with a smaller alphabet (and thus faster underlying arithmetic).

From a broader point of view, wherever Reed–Solomon codes are used, we can substitute AG codes with some benefits: either beating random constructions, or beating Reed–Solomon codes which are of bounded length for a given alphabet.

Another area of Algebraic Coding Theory with which we are more recently concerned is the one of Locally Decodable Codes. After having been first theoretically introduced, those codes now begin to find practical applications, most notably in cloud-based remote storage systems.

HYCOMES Project-Team

3. Research Program

3.1. Hybrid Systems Modeling

Systems industries today make extensive use of mathematical modeling tools to design computer controlled physical systems. This class of tools addresses the modeling of physical systems with models that are simpler than usual scientific computing problems by using only Ordinary Differential Equations (ODE) and Difference Equations but not Partial Differential Equations (PDE). This family of tools first emerged in the 1980's with SystemBuild by MatrixX (now distributed by National Instruments) followed soon by Simulink by Mathworks, with an impressive subsequent development.

In the early 90's control scientists from the University of Lund (Sweden) realized that the above approach did not support component based modeling of physical systems with reuse [0]. For instance, it was not easy to draw an electrical or hydraulic circuit by assembling component models of the various devices. The development of the Omola language by Hilding Elmqvist was a first attempt to bridge this gap by supporting some form of Differential Algebraic Equations (DAE) in the models. Modelica quickly emerged from this first attempt and became in the 2000's a major international concerted effort with the Modelica Consortium [0]. A wider set of tools, both industrial and academic, now exists in this segment [0]. In the EDA sector, VHDL-AMS was developed as a standard [12] and also allows for differential algebraic equations. Several domain-specific languages and tools for mechanical systems or electronic circuits also support some restricted classes of differential algebraic equations. Spice is the historic and most striking instance of these domain-specific languages/tools [0]. The main difference is that equations are hidden and the fixed structure of the differential algebraic results from the physical domain covered by these languages.

Despite these tools are now widely used by a number of engineers, they raise a number of technical difficulties. The meaning of some programs, their mathematical semantics, can be tainted with uncertainty. A main source of difficulty lies in the failure to properly handle the discrete and the continuous parts of systems, and their interaction. How the propagation of mode changes and resets should be handled? How to avoid artifacts due to the use of a global ODE solver causing unwanted coupling between seemingly non interacting subsystems? Also, the mixed use of an equational style for the continuous dynamics with an imperative style for the mode changes and resets is a source of difficulty when handling parallel composition. It is therefore not uncommon that tools return complex warnings for programs with many different suggested hints for fixing them. Yet, these "pathological" programs can still be executed, if wanted so, giving surprising results — See for instance the Simulink examples in [19], [15] and [16].

Indeed this area suffers from the same difficulties that led to the development of the theory of synchronous languages as an effort to fix obscure compilation schemes for discrete time equation based languages in the 1980's. Our vision is that hybrid systems modeling tools deserve similar efforts in theory as synchronous languages did for the programming of embedded systems.

3.2. Background on non-standard analysis

Non-Standard analysis plays a central role in our research on hybrid systems modeling [15], [19], [17], [16]. The following text provides a brief summary of this theory and gives some hints on its usefulness in the context of hybrid systems modeling. This presentation is based on our paper [2], a chapter of Simon Bliudze's PhD thesis [25], and a recent presentation of non-standard analysis, not axiomatic in style, due to the mathematician Lindström [49].

[0] http://www.lccc.lth.se/media/LCCC2012/WorkshopSeptember/slides/Astrom.pdf
[0] https://www.modelica.org/
[0] SimScape by Mathworks, Amesim by LMS International, now Siemens PLM, and more.
[0] http://bwrcs.eecs.berkeley.edu/Classes/IcBook/SPICE/MANUALS/spice3.html

Non-standard numbers allowed us to reconsider the semantics of hybrid systems and propose a radical alternative to the *super-dense time semantics* developed by Edward Lee and his team as part of the Ptolemy II project, where cascades of successive instants can occur in zero time by using $\mathbb{R}_+ \times \mathbb{N}$ as a time index. In the non-standard semantics, the time index is defined as a set $\mathbb{T} = \{n\partial \mid n \in {}^*\mathbb{N}\}$, where ∂ is an *infinitesimal* and ${}^*\mathbb{N}$ is the set of *non-standard integers*. Remark that (1) $\mathbb{T}$ is dense in $\mathbb{R}_+$, making it "continuous", and (2) every $t \in \mathbb{T}$ has a predecessor in $\mathbb{T}$ and a successor in $\mathbb{T}$, making it "discrete". Although it is not effective from a computability point of view, the *non-standard semantics* provides a framework that is familiar to the computer scientist and at the same time efficient as a symbolic abstraction. This makes it an excellent candidate for the development of provably correct compilation schemes and type systems for hybrid systems modeling languages.

Non-standard analysis was proposed by Abraham Robinson in the 1960s to allow the explicit manipulation of "infinitesimals" in analysis [58], [41], [11]. Robinson's approach is axiomatic; he proposes adding three new axioms to the basic Zermelo-Fraenkel (ZFC) framework. There has been much debate in the mathematical community as to whether it is worth considering non-standard analysis instead of staying with the traditional one. We do not enter this debate. The important thing for us is that non-standard analysis allows the use of the non-standard discretization of continuous dynamics "as if" it was operational.

Not surprisingly, such an idea is quite ancient. Iwasaki et al. [45] first proposed using non-standard analysis to discuss the nature of time in hybrid systems. Bliudze and Krob [26], [25] have also used non-standard analysis as a mathematical support for defining a system theory for hybrid systems. They discuss in detail the notion of "system" and investigate computability issues. The formalization they propose closely follows that of Turing machines, with a memory tape and a control mechanism.

3.3. Structural Analysis of DAE Systems

The Modelica language is based on Differential Algebraic Equations (DAE). The general form of a DAE is given by:

$$F(t, x, x', x'', \cdots) \tag{2}$$

where F is a system of n_e equations $\{f_1, \cdots, f_{n_e}\}$ and x is a finite list of n_v independent real-valued, smooth enough, functions $\{x_1, \cdots, x_{n_v}\}$ of the independent variable t. We use x' as a shorthand for the list of first-order time derivatives of x_j, $j = 1, \cdots, n_v$. High-order derivatives are recursively defined as usual, and $x^{(k)}$ denotes the list formed by the k-th derivatives of the functions x_j. Each f_i depends on the scalar t and some of the functions x_j as well as a finite number of their derivatives.

Let $\sigma_{i,j}$ denote the highest differentiation order of variable x_j effectively appearing in equation f_i, or $-\infty$ if x_j does not appear in f_i. The *leading variables* of F are the variables in the set

$$\left\{x_j^{(\sigma_j)} \mid \sigma_j = \max_i \sigma_{i,j}\right\}$$

The *state variables* of F are the variables in the set

$$\left\{x_j^{(\nu_j)} \mid 0 \leq \nu_j < \max_i \sigma_{i,j}\right\}$$

A leading variable $x_j^{(\sigma_j)}$ is said to be *algebraic* if $\sigma_j = 0$ (in which case, neither x_j nor any of its derivatives are state variables). In the sequel, v and u denote the leading and state variables of F, respectively.

DAE are a strict generalization of *ordinary differential equations* (*ODE*), in the sense that it may not be immediate to rewrite a DAE as an explicit ODE of the form $v = G(u)$. The reason is that this transformation relies on the Implicit Function Theorem, requiring that the Jacobian matrix $\frac{\partial F}{\partial v}$ have full rank. This is, in general, not the case for a DAE. Simple examples, like the two-dimensional fixed-length pendulum in Cartesian coordinates [55], exhibit this behaviour.

For a square DAE of dimension n (i.e., we now assume $n_e = n_v = n$) to be solved in the neighborhood of some (v^*, u^*), one needs to find a set of non-negative integers $C = \{c_1, \cdots, c_n\}$ such that system

$$F^{(C)} = \{f_1^{(c_1)}, \cdots, f_n^{(c_n)}\}$$

can locally be made explicit, i.e., the Jacobian matrix of $F^{(C)}$ with respect to its leading variables, evaluated at (v^*, u^*), is nonsingular. The smallest possible value of $\max_i c_i$ for a set C that satisfies this property is the *differentiation index* [32] of F, that is, the minimal number of time differentiations of all or part of the equations f_i required to get an ODE.

In practice, the problem of automatically finding a "minimal" solution C to this problem quickly becomes intractable. Moreover, the differentiation index may depend on the value of (v^*, u^*). This is why, in lieu of numerical nonsingularity, one is interested in the *structural nonsingularity* of the Jacobian matrix, i.e., its almost certain nonsingularity when its nonzero entries vary over some neighborhood. In this framework, the *structural analysis* (*SA*) of a DAE returns, when successful, values of the c_i that are independent from a given value of (v^*, u^*).

A renowned method for the SA of DAE is the *Pantelides method*; however, Pryce's Σ*-method* is introduced also in what follows, as it is a crucial tool for our works.

3.3.1. Pantelides method

In 1988, Pantelides proposed what is probably the most well-known SA method for DAE [55]. The leading idea of his work is that the structural representation of a DAE can be condensed into a bipartite graph whose left nodes (resp. right nodes) represent the equations (resp. the variables), and in which an edge exists if and only if the variable occurs in the equation.

By detecting specific subsets of the nodes, called *Minimally Structurally Singular* (*MSS*) subsets, the Pantelides method iteratively differentiates part of the equations until a perfect matching between the equations and the leading variables is found. One can easily prove that this is a necessary and sufficient condition for the structural nonsingularity of the system.

The main reason why the Pantelides method is not used in our work is that it cannot efficiently be adapted to multimode DAE (mDAE). As a matter of fact, the adjacency graph of a mDAE has both its nodes and edges parametrized by the subset of modes in which they are active; this, in turn, requires that a parametrized Pantelides method must branch every time no mode-independent MSS is found, ultimately resulting, in the worst case, in the enumeration of modes.

3.3.2. Pryce's Σ-method

Albeit less renowned that the Pantelides method, Pryce's Σ-method [56] is an efficient SA method for DAE, whose equivalence to the Pantelides method has been proved by the author. This method consists in solving two successive problems, denoted by primal and dual, relying on the Σ*-matrix*, or *signature matrix*, of the DAE F.

This matrix is given by:

$$\Sigma = (\sigma_{ij})_{1 \leq i,j \leq n} \tag{3}$$

where σ_{ij} is equal to the greatest integer k such that $x_j^{(k)}$ appears in f_i, or $-\infty$ if variable x_j does not appear in f_i. It is the adjacency matrix of a weighted bipartite graph, with structure similar to the graph considered in the Pantelides method, but whose edges are weighted by the highest differentiation orders. The $-\infty$ entries denote non-existent edges.

The *primal problem* consists in finding a *maximum-weight perfect matching* (*MWPM*) in the weighted adjacency graph. This is actually an assignment problem, for the solving of which several standard algorithms exist, such as the push-relabel algorithm [44] or the Edmonds-Karp algorithm [43] to only give a few. However, none of these algorithms are easily parametrizable, even for applications to mDAE systems with a fixed number of variables.

The *dual problem* consists in finding the component-wise minimal solution $(C, D) = (\{c_1, \cdots, c_n\}, \{d_1, \cdots, d_n\})$ to a given linear programming problem, defined as the dual of the aforementioned assignment problem. This is performed by means of a *fixpoint iteration* (*FPI*) that makes use of the MWPM found as a solution to the primal problem, described by the set of tuples $\{(i, j_i)\}_{i\in\{1,\cdots,n\}}$:

1. Initialize $\{c_1, \cdots, c_n\}$ to the zero vector.
2. For every $j \in \{1, \cdots, n\}$,
$$d_j \longleftarrow \max_i(\sigma_{ij} + c_i)$$
3. For every $i \in \{1, \cdots, n\}$,
$$c_i \longleftarrow d_{j_i} - \sigma_{i,j_i}$$
4. Repeat Steps 2 and 3 until convergence is reached.

From the results proved by Pryce in [56], it is known that the above algorithm terminates if and only if it is provided a MWPM, and that the values it returns are independent of the choice of a MWPM whenever there exist several such matchings. In particular, a direct corollary is that the Σ-method succeeds as long as a perfect matching can be found between equations and variables.

Another important result is that, if the Pantelides method succeeds for a given DAE F, then the Σ-method also succeeds for F and the values it returns for C are exactly the differentiation indices for the equations that are returned by the Pantelides method. As for the values of the d_j, being given by $d_j = \max_i(\sigma_{ij} + c_i)$, they are the differentiation indices of the leading variables in $F^{(C)}$.

Working with this method is natural for our works, since the algorithm for solving the dual problem is easily parametrizable for dealing with multimode systems, as shown in our recent paper [31].

3.3.3. Block triangular decomposition

Once structural analysis has been performed, system $F^{(C)}$ can be regarded, for the needs of numerical solving, as an algebraic system with unknowns $x_j^{(d_j)}$, $j = 1 \cdots n$. As such, (inter)dependencies between its equations must be taken into account in order to put it into block triangular form (BTF). Three steps are required:

1. the *dependency graph* of system $F^{(C)}$ is generated, by taking into account the perfect matching between equations $f_i^{(c_i)}$ and unknowns $x_j^{(d_j)}$;
2. the *strongly connected components* (*SCC*) in this graph are determined: these will be the *equation blocks* that have to be solved;
3. the *block dependency graph* is constructed as the condensation of the dependency graph, from the knowledge of the SCC; a BTF of system $F^{(C)}$ can be made explicit from this graph.

3.4. Contract-Based Design, Interfaces Theories, and Requirements Engineering

System companies such as automotive and aeronautic companies are facing significant difficulties due to the exponentially raising complexity of their products coupled with increasingly tight demands on functionality, correctness, and time-to-market. The cost of being late to market or of imperfections in the products is staggering as witnessed by the recent recalls and delivery delays that many major car and airplane manufacturers had to bear in the recent years. The specific root causes of these design problems are complex and relate to a number of issues ranging from design processes and relationships with different departments of the same company and with suppliers, to incomplete requirement specification and testing.

We believe the most promising means to address the challenges in systems engineering is to employ structured and formal design methodologies that seamlessly and coherently combine the various viewpoints of the design space (behavior, space, time, energy, reliability, ...), that provide the appropriate abstractions to manage the inherent complexity, and that can provide correct-by-construction implementations. The following technology issues must be addressed when developing new approaches to the design of complex systems:

- The overall design flows for heterogeneous systems and the associated use of models across traditional boundaries are not well developed and understood. Relationships between different teams inside a same company, or between different stake-holders in the supplier chain, are not well supported by solid technical descriptions for the mutual obligations.
- System requirements capture and analysis is in large part a heuristic process, where the informal text and natural language-based techniques in use today are facing significant challenges [10]. Formal requirements engineering is in its infancy: mathematical models, formal analysis techniques and links to system implementation must be developed.
- Dealing with variability, uncertainty, and life-cycle issues, such as extensibility of a product family, are not well-addressed using available systems engineering methodologies and tools.

The challenge is to address the entire process and not to consider only local solutions of methodology, tools, and models that ease part of the design.

Contract-based design has been proposed as a new approach to the system design problem that is rigorous and effective in dealing with the problems and challenges described before, and that, at the same time, does not require a radical change in the way industrial designers carry out their task as it cuts across design flows of different type. Indeed, contracts can be used almost everywhere and at nearly all stages of system design, from early requirements capture, to embedded computing infrastructure and detailed design involving circuits and other hardware. Contracts explicitly handle pairs of properties, respectively representing the assumptions on the environment and the guarantees of the system under these assumptions. Intuitively, a contract is a pair $C = (A, G)$ of assumptions and guarantees characterizing in a formal way 1) under which context the design is assumed to operate, and 2) what its obligations are. Assume/Guarantee reasoning has been known for a long time, and has been used mostly as verification mean for the design of software [53]. However, contract based design with explicit assumptions is a philosophy that should be followed all along the design, with all kinds of models, whenever necessary. Here, specifications are not limited to profiles, types, or taxonomy of data, but also describe the functions, performances of various kinds (time and energy), and reliability. This amounts to enrich a component's interface with, on one hand, formal specifications of the behavior of the environment in which the component may be instantiated and, on the other hand, of the expected behavior of the component itself. The consideration of rich interfaces is still in its infancy. So far, academic researchers have addressed the mathematics and algorithmics of interfaces theories and contract-based reasoning. To make them a technique of choice for system engineers, we must develop:

- Mathematical foundations for interfaces and requirements engineering that enable the design of frameworks and tools;
- A system engineering framework and associated methodologies and tool sets that focus on system requirements modeling, contract specification, and verification at multiple abstraction layers.

A detailed bibliography on contract and interface theories for embedded system design can be found in [3]. In a nutshell, contract and interface theories fall into two main categories:

Assume/guarantee contracts. By explicitly relying on the notions of assumptions and guarantees, A/G-contracts are intuitive, which makes them appealing for the engineer. In A/G-contracts, assumptions and guarantees are just properties regarding the behavior of a component and of its environment. The typical case is when these properties are formal languages or sets of traces, which includes the class of safety properties [46], [35], [52], [14], [37]. Contract theories were initially developed as specification formalisms able to refuse some inputs from the environment [42]. A/G-contracts were advocated in [18] and are is still a very active research topic, with several contributions dealing with the timed [24] and probabilistic [29], [30] viewpoints in system design, and even mixed-analog circuit design [54].

Automata theoretic interfaces. Interfaces combine assumptions and guarantees in a single, automata theoretic specification. Most interface theories are based on Lynch Input/Output Automata [51], [50]. Interface Automata [61], [60], [62], [33] focus primarily on parallel composition and compatibility: Two interfaces can be composed and are compatible if there is at least one environment where they can work together. The idea is that the resulting composition exposes as an interface the needed information to ensure that incompatible pairs of states cannot be reached. This can be achieved by using the possibility, for an Interface Automaton, to refuse selected inputs from the environment in a given state, which amounts to the implicit assumption that the environment will never produce any of the refused inputs, when the interface is in this state. Modal Interfaces [57] inherit from both Interface Automata and the originally unrelated notion of Modal Transition System [48], [13], [27], [47]. Modal Interfaces are strictly more expressive than Interface Automata by decoupling the I/O orientation of an event and its deontic modalities (mandatory, allowed or forbidden). Informally, a *must* transition is available in every component that realizes the modal interface, while a *may* transition needs not be. Research on interface theories is still very active. For instance, timed [63], [21], [23], [39], [38], [22], probabilistic [29], [40] and energy-aware [34] interface theories have been proposed recently.

Requirements Engineering is one of the major concerns in large systems industries today, particularly so in sectors where certification prevails [59]. Most requirements engineering tools offer a poor structuring of the requirements and cannot be considered as formal modeling frameworks today. They are nothing less, but nothing more than an informal structured documentation enriched with hyperlinks. As examples, medium size sub-systems may have a few thousands requirements and the Rafale fighter aircraft has above 250,000 of them. For the Boeing 787, requirements were not stable while subcontractors were working on the development of the fly-by-wire and of the landing gear subsystems, leading to a long and cahotic convergence of the design process.

We see Contract-Based Design and Interfaces Theories as innovative tools in support of Requirements Engineering. The Software Engineering community has extensively covered several aspects of Requirements Engineering, in particular:

- the development and use of large and rich *ontologies*; and
- the use of Model Driven Engineering technology for the structural aspects of requirements and resulting hyperlinks (to tests, documentation, PLM, architecture, and so on).

Behavioral models and properties, however, are not properly encompassed by the above approaches. This is the cause of a remaining gap between this phase of systems design and later phases where formal model based methods involving behavior have become prevalent—see the success of Matlab/Simulink/Scade technologies. We believe that our work on contract based design and interface theories is best suited to bridge this gap.

Kairos Project-Team

3. Research Program

3.1. Cyber-Physical co-modeling

Cyber-Physical System modeling requires joint representation of digital/cyber controllers and natural physics environments. Heterogeneous modeling must then be articulated to support accurate (co-)simulation, (co-)analysis, and (co-)verification. The picture above sketches the overall design framework. It comprises functional requirements, to be met provided surrounding platform guarantees, in a contract approach. All relevant aspects are modeled with proper Domain Specific Languages (DSL), so that constraints can be gathered globally, then analyzed to build a mapping proposal with both a structural aspect (functions allocated to platform resources), but also a behavioral ones, scheduling activities. Mapping may be computed automatically or not, provably correct or not, obtained by static analytic methods or abstract execution. Physical phenomena (in a very broad acceptance of the term) are usually modeled using continuous-time models and differential equations. Then the "proper" discretization opportunities for numerical simulation form a large spectrum of mathematical engineering practices. This is not at all the domain of expertise of Kairos members, but it should not be a limitation as long as one can assume a number of properties from the discretized version. On the other hand, we do have a strong expertise on modeling of both embedded processing architectures and embedded software (i.e., the kind of usually concurrent, sometimes distributed software that reacts to and control the physical environment). This is important as, unlike in the "physical" areas where modeling is common-place, modeling of software and programs is far from mainstream in the Software Engineering community. These domains are also an area of computer science where modeling, and even formal modeling, of the real objects that are originally of discrete/cyber nature, takes some importance with formal Models of Computation and Communications. It seems therefore quite natural to combine physical and cyber modeling in a more global design approach (even multi-physic domains and systems of systems possibly, but always with software-intensive aspects involved). Our objective is certainly not to become experts in physical modeling and/or simulation process, but to retain from it only the essential and important aspects to include them into System-Level Engineering design, based on Model-Driven approaches allowing formal analysis.

This sets an original research agenda: Model-Based System Engineering environments exist, at various stages of maturity and specificity, in the academic and industrial worlds. Formal Methods and Verification/Certification techniques also exist, but generally in a point-wise fashion. Our approach aims at raising the level of formality describing relevant features of existing individual models, so that formal methods can have a greater general impact on usual, "industrial-level", modeling practices. Meanwhile, the relevance of formal methods is enhanced as it now covers various aspects in a uniform setting (timeliness, energy budget, dependability, safety/security...).

New research directions on formal CPS design should focus on the introduction of uncertainty (stochastic models) in our particular framework, on relations between (logical) real-time and security, on relations between common programming languages paradigms and logical time, on extending logical frameworks with logical time, on the concern with resource discovery also in presence of mobility inherent to connected objects and Internet of Things.

3.2. Cyber-Physical co-simulation

The FMI standard (Functional Mock-Up Interface) has been proposed for "purely physical" (i.e., based on persistent signals) co-simulation, and then adopted in over 100 industrial tools including frameworks such as Matlab/Simulink and Ansys, to mention two famous model editors. With the recent use of co-simulation to cyber-physical systems, dealing with the discrete and transient nature of cyber systems became mandatory. Together with other people from the community, we shown that FMI and other frameworks for co-simulation

badly support co-simulation of cyber-physical systems; leading to bad accuracy and performances. More precisely, the way to interact with the different parts of the co-simulation require a specific knowledge about its internal semantics and the kind of data exposed (e.g., continuous, piecewise-constant). Towards a better co-simulation of cyber-physical systems, we are looking for conservative abstractions of the parts and formalisms that aim to describe the functional and temporal constraints that are required to bind several simulation models together.

3.3. Formal analysis and verification

Because the nature of our constraints is specific, we want to adjust verification methods to the goals and expressiveness of our modeling approach. Quantitative (interval) timing conditions on physical models combined with (discrete) cyber modes suggest the use of SMT (Satisfiability Modulo Theories) automatic solvers, but the natural expressiveness requested (as for instance in our CCSL constructs) shows this is not always feasible. Either interactive proofs, or suboptimal solutions (essentially resulting of abstract run-time simulations) should be considered. Complementarily to these approaches, we are experimenting with new variants of symbolic behavioural semantics, allowing to construct finite representations of the behaviour of CPS systems with explicit handling of data, time, or other non-functional aspects.

3.4. Relation between Model and Code

While models considered in Kairos can also be considered as executable specifications (through abstract simulation schemes), they can also lead to code synthesis and deployment. Conversely, code execution of smaller, elementary software components can lead to performance estimation enriching the models before global mapping optimization. CPS introduce new challenging problems for code performance stability. Indeed, two additional factors for performance variability appear, which were not present in classical embedded systems: 1) variable and continuous data input from the physical world and 2) variable underlying hardware platform. For the first factor, CPS software must be analysed in conjunction with its data input coming from the physics, so the variability of the performance may come from the various data. For the second factor, the underlying hardware of the CPS may change during the time (new computing actors appear or disappear, some actors can be reconfigured during execution). The new challenge is to understand how these factors influence performance variability exactly, and how to provide solutions to reduce it or to model it. The modeling of performance variability becomes a new input.

3.5. Code generation and optimization

A significant part CPS design happens at model level, through activities such as model construction, analysis, or verification. However, in most cases the objective of the design process is implementation. We mostly consider the implementation problem in the context of embedded, real-time, or edge computing applications, which are subject to stringent performance, embedding, and safety *non-functional requirements*.

The implementation of such systems usually involves a mix of synthesis—(real-time) scheduling, code generation, compilation—and performance (*e.g.* timing) analysis. One key difficulty here is that synthesis and performance analysis depend on each other. As enumerating the various solutions is not possible for complexity reasons, heuristic implementation methods are needed in all cases. One popular solution here is to build the system first using unsafe performance estimations for its components, and then check system *schedulability* through a global analysis. Another solution is to use safe, over-approximated performance estimations and perform their mapping in a way that ensures by construction the schedulability of the system.

In both cases, the specification of the design space—functional specification, execution platform model, non-functional requirements, implementation model—is a key problem. Another problem is the definition of scalable and efficient mapping methods based on both "exact" approaches (ILP/SMT/CP solving) and compilation-like heuristics.

3.6. Extending logical frameworks with logical time

The Curry-Howard isomorphism (*proposition-as-types and proofs-as-typed-λ-terms*) represent the logical and computational basis to interactive theorem provers: our challenge is to investigate and design time constraints within a Dependent Type Theory (e.g. if event A happened-before event B, then the timestamp/type of A is less (i.e. a subtype) than the timestamp/type of B). We are currently extending the Edinburgh Logical Framework (LF) of Harper-Honsell-Plotkin with relevant constructs expressing logical time and synchronization between processes. Also, union and intersection types with their subtyping constraints theories could capture some constraints expressions *à la* CCSL needed to formalize logical clocks (in particular CCSL expressions like subclock, clock union, intersection and concatenation) and provide opportunities for an *ad hoc* polymorphic timed Type Theory. Logical time constraints seen as property types can beneficially be handled by logical frameworks. The new challenge here is to demonstrate the relevance of Type Theory to work on logical and multiform timing constraint resolution.

3.7. Object-oriented programming and logical time

We formalize in the past object-oriented programming features and safe static type systems featuring delegation-based or trait inheritance: well typed program will never produce into the `message-not-found` infamous run-time error. We view the logical time as a means to enhance the description of timing constraints and properties on top of existing language semantics. When considering general purpose object-oriented languages, like Java, Type Theory is a natural way to provide such properties. Currently, few languages have special types to manage instants, time structures and instant relations like subclocking, precedence, causality, equality, coincidence, exclusion, independence, etc. CCSL provides ad-hoc constructors to specify clock constraints and logical time: enriching object-oriented type theories with CCSL expressions could constitute an interesting research perspective towards a wider usage of CCSL. The new challenge is to consider logical time constraints as behavioral type properties, and the design of programming language constructs and *ad-hoc* type systems. Advances of typed-calculi featuring those static time features will be applied to our extension [42] of the lambda-calculus of objects of Fisher-Honsell-Mitchell.

3.8. Extensions for spatio-temporal modeling and mobile systems

While Time is clearly a primary ingredient in the proper design of CPS systems, in some cases Space, and related notions of local proximity or conversely long distance, play also a key role for correct modeling, often in part because of the constraints this puts on interactions and time for communications. Once space is taken into account, one has to recognize also that many systems will request to consider mobility, originated as change of location through time. Mobile CPS (or mCPS) systems occur casually, e.g., in the case of Intelligent Transportation Systems, or in roaming connected objects of the IoT. Spatio-temporal and mobility modeling may each lead to dynamicity in the representation of constraints, with the creation/deletion/discovering of new components in the system. This opportunity for new expressiveness will certainly cause new needs in handling constraint systems and topological graph locations. The new challenge is to provide an algebraic support with a constraint description language that could be as simple and expressive as possible, and of use in the semantic annotations for mobile CPS design. We also aims to provide fully distributed routing protocols to manage Semantic Resource Discovery in IoT.

KOPERNIC Team

3. Research Program

3.1. Worst case execution time estimation of a program

Modern processors induce an increased variability of the execution time of programs, making difficult (or even impossible) a complete static analysis. Our objective is to propose a solution composing probabilistic and non-probabilistic approaches based both on static and on statistical analyses by answering the following **scientific challenges**:

1. **a classification of the variability of execution times** of a program with respect to the processor features. We will use as first measure our statistical estimator based on the Extreme Value Theory [18], [20]. An implementation of the estimator is available at http://inria-rscript.serveftp.com. The access to this later page requires a login (aoste) and a password (aoste). The difficulty of this challenge is related to the definition of an element belonging to the set of variability factors and its mapping to the execution time of the program.
2. **a compositional rule** of statistical models based on Bayesian approaches. The difficulty of this challenge comes from the fact that a global maximum cannot be obtained by upper bounding the corresponding local maxima. We will use as first rule of composition a Bayesian approach [22]. We consider as first statistical model those obtained by any static analysis of the program on a basic processor. Through the Bayesian approach we add iteratively the variability due to each processor feature as a new statistical model. The convergence of the global model is decided once no variability is detected at the level of the statistical estimator providing the bounds on the execution time of the program.

The problem of estimating the worst case execution time of a program is an excellent opportunity for the Extreme Values community to validate and to evolve as the context of obtaining measures is indefinitely reproducible.

3.2. Deciding the schedulability of all programs running within the same cyber component

In this context, the programs may have different time criticalities, but they share the same processor, possibly multicore [0]. Our objective is to propose a solution composing probabilistic and non-probabilistic approaches based on answers to the following **scientific challenges**:

1. **scheduling algorithms taking into account the interaction between different variability factors**. The proposed scheduling algorithms are the theoretical bases of a scheduler able to guarantee the time constraints of the cyber component. The existence of time parameters described by probability distributions imposes to answer to the challenge of revisiting scheduling algorithms that lose their optimality even in the case of an unicore processor [26]. Moreover, the multicore partionning problem is, also, recognized difficult for the non-probabilistic case [29];
2. **schedulability analyses** based on the algorithms proposed previously. In the case of predictable processors, the schedulability analyses accounting for operating systems costs increase the dependability of CPSs [28]. Morever, in presence of variability factors, the additivity property of non-probabilistic approaches is lost and new composition principles are required. We will propose new composition principles based on our preliminary results on the propagation of the probabilistic constraints [16]. The definition of these principles form the challenge related to this objective.

[0] This case is referred as a mixed criticality approach.

3.3. Deciding the schedulability of all programs communicating through predictable and non-predictable networks

In this case the programs of the same cyber component execute on the same processor and they may communicate with the programs of other cyber components through networks that may be predictable (network on chip) or non-predictable (internet, telecommunications). Our objective is to propose a solution to the challenge of analysing schedulability of programs, for which existing (worst case) probabilistic solutions exist [27], communicating through networks, for which probabilistic worst-case solutions [19] and average solutions exist [24]. Our solution is based on the results obtained for the two first objectives, making this third objective a longer-term one.

LFANT Project-Team

3. Research Program

3.1. Number fields, class groups and other invariants

Participants: Bill Allombert, Jared Guissmo Asuncion, Karim Belabas, Jean-Paul Cerri, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Fredrik Johansson, Aurel Page.

Modern number theory has been introduced in the second half of the 19th century by Dedekind, Kummer, Kronecker, Weber and others, motivated by Fermat's conjecture: There is no non-trivial solution in integers to the equation $x^n + y^n = z^n$ for $n \geqslant 3$. Kummer's idea for solving Fermat's problem was to rewrite the equation as $(x+y)(x+\zeta y)(x+\zeta^2 y)\cdots(x+\zeta^{n-1}y) = z^n$ for a primitive n-th root of unity ζ, which seems to imply that each factor on the left hand side is an n-th power, from which a contradiction can be derived.

The solution requires to augment the integers by *algebraic numbers*, that are roots of polynomials in $\mathbb{Z}[X]$. For instance, ζ is a root of $X^n - 1$, $\sqrt[3]{2}$ is a root of $X^3 - 2$ and $\frac{\sqrt{3}}{5}$ is a root of $25X^2 - 3$. A *number field* consists of the rationals to which have been added finitely many algebraic numbers together with their sums, differences, products and quotients. It turns out that actually one generator suffices, and any number field K is isomorphic to $\mathbb{Q}[X]/(f(X))$, where $f(X)$ is the minimal polynomial of the generator. Of special interest are *algebraic integers*, "numbers without denominators", that are roots of a monic polynomial. For instance, ζ and $\sqrt[3]{2}$ are integers, while $\frac{\sqrt{3}}{5}$ is not. The *ring of integers* of K is denoted by $\mathcal{O}_K$; it plays the same role in K as $\mathbb{Z}$ in $\mathbb{Q}$.

Unfortunately, elements in $\mathcal{O}_K$ may factor in different ways, which invalidates Kummer's argumentation. Unique factorisation may be recovered by switching to *ideals*, subsets of $\mathcal{O}_K$ that are closed under addition and under multiplication by elements of $\mathcal{O}_K$. In $\mathbb{Z}$, for instance, any ideal is *principal*, that is, generated by one element, so that ideals and numbers are essentially the same. In particular, the unique factorisation of ideals then implies the unique factorisation of numbers. In general, this is not the case, and the *class group* Cl_K of ideals of $\mathcal{O}_K$ modulo principal ideals and its *class number* $h_K = |\,\mathrm{Cl}_K\,|$ measure how far $\mathcal{O}_K$ is from behaving like $\mathbb{Z}$.

Using ideals introduces the additional difficulty of having to deal with *units*, the invertible elements of $\mathcal{O}_K$: Even when $h_K = 1$, a factorisation of ideals does not immediately yield a factorisation of numbers, since ideal generators are only defined up to units. For instance, the ideal factorisation $(6) = (2)\cdot(3)$ corresponds to the two factorisations $6 = 2\cdot 3$ and $6 = (-2)\cdot(-3)$. While in $\mathbb{Z}$, the only units are 1 and -1, the unit structure in general is that of a finitely generated $\mathbb{Z}$-module, whose generators are the *fundamental units*. The *regulator* R_K measures the "size" of the fundamental units as the volume of an associated lattice.

One of the main concerns of algorithmic algebraic number theory is to explicitly compute these invariants (Cl_K and h_K, fundamental units and R_K), as well as to provide the data allowing to efficiently compute with numbers and ideals of $\mathcal{O}_K$; see [36] for a recent account.

The *analytic class number formula* links the invariants h_K and R_K (unfortunately, only their product) to the ζ-function of K, $\zeta_K(s) := \prod_{\mathfrak{p} \text{ prime ideal of } \mathcal{O}_K} \left(1 - \mathrm{N}\,\mathfrak{p}^{-s}\right)^{-1}$, which is meaningful when $\Re(s) > 1$, but which may be extended to arbitrary complex $s \neq 1$. Introducing characters on the class group yields a generalisation of ζ- to L-functions. The *generalised Riemann hypothesis (GRH)*, which remains unproved even over the rationals, states that any such L-function does not vanish in the right half-plane $\Re(s) > 1/2$. The validity of the GRH has a dramatic impact on the performance of number theoretic algorithms. For instance, under GRH, the class group admits a system of generators of polynomial size; without GRH, only exponential bounds are known. Consequently, an algorithm to compute Cl_K via generators and relations (currently the only viable practical approach) either has to assume that GRH is true or immediately becomes exponential.

When $h_K = 1$ the number field K may be norm-Euclidean, endowing $\mathcal{O}_K$ with a Euclidean division algorithm. This question leads to the notions of the Euclidean minimum and spectrum of K, and another task in algorithmic number theory is to compute explicitly this minimum and the upper part of this spectrum, yielding for instance generalised Euclidean gcd algorithms.

3.2. Function fields, algebraic curves and cryptology

Participants: Karim Belabas, Guilhem Castagnos, Jean-Marc Couveignes, Andreas Enge, Damien Robert, Jean Kieffer, Razvan Barbulescu.

Algebraic curves over finite fields are used to build the currently most competitive public key cryptosystems. Such a curve is given by a bivariate equation $\mathcal{C}(X, Y) = 0$ with coefficients in a finite field $\mathbb{F}_q$. The main classes of curves that are interesting from a cryptographic perspective are *elliptic curves* of equation $\mathcal{C} = Y^2 - (X^3 + aX + b)$ and *hyperelliptic curves* of equation $\mathcal{C} = Y^2 - (X^{2g+1} + \cdots)$ with $g \geqslant 2$.

The cryptosystem is implemented in an associated finite abelian group, the *Jacobian* $\mathrm{Jac}_\mathcal{C}$. Using the language of function fields exhibits a close analogy to the number fields discussed in the previous section. Let $\mathbb{F}_q(X)$ (the analogue of $\mathbb{Q}$) be the *rational function field* with subring $\mathbb{F}_q[X]$ (which is principal just as $\mathbb{Z}$). The *function field* of $\mathcal{C}$ is $K_\mathcal{C} = \mathbb{F}_q(X)[Y]/(\mathcal{C})$; it contains the *coordinate ring* $\mathcal{O}_\mathcal{C} = \mathbb{F}_q[X, Y]/(\mathcal{C})$. Definitions and properties carry over from the number field case $K/\mathbb{Q}$ to the function field extension $K_\mathcal{C}/\mathbb{F}_q(X)$. The Jacobian $\mathrm{Jac}_\mathcal{C}$ is the divisor class group of $K_\mathcal{C}$, which is an extension of (and for the curves used in cryptography usually equals) the ideal class group of $\mathcal{O}_\mathcal{C}$.

The size of the Jacobian group, the main security parameter of the cryptosystem, is given by an L-function. The GRH for function fields, which has been proved by Weil, yields the Hasse–Weil bound $(\sqrt{q} - 1)^{2g} \leqslant |\mathrm{Jac}_\mathcal{C}| \leqslant (\sqrt{q} + 1)^{2g}$, or $|\mathrm{Jac}_\mathcal{C}| \approx q^g$, where the *genus* g is an invariant of the curve that correlates with the degree of its equation. For instance, the genus of an elliptic curve is 1, that of a hyperelliptic one is $\frac{\deg_X \mathcal{C} - 1}{2}$. An important algorithmic question is to compute the exact cardinality of the Jacobian.

The security of the cryptosystem requires more precisely that the *discrete logarithm problem* (DLP) be difficult in the underlying group; that is, given elements D_1 and $D_2 = xD_1$ of $\mathrm{Jac}_\mathcal{C}$, it must be difficult to determine x. Computing x corresponds in fact to computing $\mathrm{Jac}_\mathcal{C}$ explicitly with an isomorphism to an abstract product of finite cyclic groups; in this sense, the DLP amounts to computing the class group in the function field setting.

For any integer n, the *Weil pairing* e_n on $\mathcal{C}$ is a function that takes as input two elements of order n of $\mathrm{Jac}_\mathcal{C}$ and maps them into the multiplicative group of a finite field extension $\mathbb{F}_{q^k}$ with $k = k(n)$ depending on n. It is bilinear in both its arguments, which allows to transport the DLP from a curve into a finite field, where it is potentially easier to solve. The *Tate-Lichtenbaum pairing*, that is more difficult to define, but more efficient to implement, has similar properties. From a constructive point of view, the last few years have seen a wealth of cryptosystems with attractive novel properties relying on pairings.

For a random curve, the parameter k usually becomes so big that the result of a pairing cannot even be output any more. One of the major algorithmic problems related to pairings is thus the construction of curves with a given, smallish k.

3.3. Complex multiplication

Participants: Jared Guissmo Asuncion, Karim Belabas, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Fredrik Johansson, Chloe Martindale, Damien Robert.

Complex multiplication provides a link between number fields and algebraic curves; for a concise introduction in the elliptic curve case, see [38], for more background material, [37]. In fact, for most curves $\mathcal{C}$ over a finite field, the endomorphism ring of $\mathrm{Jac}_\mathcal{C}$, which determines its L-function and thus its cardinality, is an order in a special kind of number field K, called *CM field*. The CM field of an elliptic curve is an imaginary-quadratic field $\mathbb{Q}(\sqrt{D})$ with $D < 0$, that of a hyperelliptic curve of genus g is an imaginary-quadratic extension of a totally real number field of degree g. Deuring's lifting theorem ensures that $\mathcal{C}$ is the reduction modulo some prime of a curve with the same endomorphism ring, but defined over the *Hilbert class field* H_K of K.

Algebraically, H_K is defined as the maximal unramified abelian extension of K; the Galois group of H_K/K is then precisely the class group Cl_K. A number field extension H/K is called *Galois* if $H \simeq K[X]/(f)$ and H contains all complex roots of f. For instance, $\mathbb{Q}(\sqrt{2})$ is Galois since it contains not only $\sqrt{2}$, but also the second root $-\sqrt{2}$ of X^2-2, whereas $\mathbb{Q}(\sqrt[3]{2})$ is not Galois, since it does not contain the root $e^{2\pi i/3}\sqrt[3]{2}$ of X^3-2. The *Galois group* $\mathrm{Gal}_{H/K}$ is the group of automorphisms of H that fix K; it permutes the roots of f. Finally, an *abelian* extension is a Galois extension with abelian Galois group.

Analytically, in the elliptic case H_K may be obtained by adjoining to K the *singular value* $j(\tau)$ for a complex valued, so-called *modular* function j in some $\tau \in \mathcal{O}_K$; the correspondence between $\mathrm{Gal}_{H/K}$ and Cl_K allows to obtain the different roots of the minimal polynomial f of $j(\tau)$ and finally f itself. A similar, more involved construction can be used for hyperelliptic curves. This direct application of complex multiplication yields algebraic curves whose L-functions are known beforehand; in particular, it is the only possible way of obtaining ordinary curves for pairing-based cryptosystems.

The same theory can be used to develop algorithms that, given an arbitrary curve over a finite field, compute its L-function.

A generalisation is provided by *ray class fields*; these are still abelian, but allow for some well-controlled ramification. The tools for explicitly constructing such class fields are similar to those used for Hilbert class fields.

MEXICO Project-Team

3. Research Program

3.1. Concurrency

Participants: Thomas Chatain, Philippe Dague, Stefan Haar, Serge Haddad, Stefan Schwoon.

Concurrency; Semantics; Automatic Control ; Diagnosis ; Verification

Concurrency: Property of systems allowing some interacting processes to be executed in parallel.

Diagnosis: The process of deducing from a partial observation of a system aspects of the internal states or events of that system; in particular, *fault diagnosis* aims at determining whether or not some non-observable fault event has occurred.

Conformance Testing: Feeding dedicated input into an implemented system IS and deducing, from the resulting output of I, whether I respects a formal specification S.

3.1.1. Introduction

It is well known that, whatever the intended form of analysis or control, a *global* view of the system state leads to overwhelming numbers of states and transitions, thus slowing down algorithms that need to explore the state space. Worse yet, it often blurs the mechanics that are at work rather than exhibiting them. Conversely, respecting concurrency relations avoids exhaustive enumeration of interleavings. It allows us to focus on 'essential' properties of non-sequential processes, which are expressible with causal precedence relations. These precedence relations are usually called causal (partial) orders. Concurrency is the explicit absence of such a precedence between actions that do not have to wait for one another. Both causal orders and concurrency are in fact essential elements of a specification. This is especially true when the specification is constructed in a distributed and modular way. Making these ordering relations explicit requires to leave the framework of state/interleaving based semantics. Therefore, we need to develop new dedicated algorithms for tasks such as conformance testing, fault diagnosis, or control for distributed discrete systems. Existing solutions for these problems often rely on centralized sequential models which do not scale up well.

3.1.2. Diagnosis

Participants: Stefan Haar, Serge Haddad, Stefan Schwoon, Philippe Dague, Lina Ye.

Fault Diagnosis for discrete event systems is a crucial task in automatic control. Our focus is on *event oriented* (as opposed to *state oriented*) model-based diagnosis, asking e.g. the following questions:
given a - potentially large - *alarm pattern* formed of observations,

- what are the possible *fault scenarios* in the system that *explain* the pattern ?
- Based on the observations, can we deduce whether or not a certain - invisible - fault has actually occurred ?

Model-based diagnosis starts from a discrete event model of the observed system - or rather, its relevant aspects, such as possible fault propagations, abstracting away other dimensions. From this model, an extraction or unfolding process, guided by the observation, produces recursively the explanation candidates.

In asynchronous partial-order based diagnosis with Petri nets [45], [46], [47], one unfolds the *labelled product* of a Petri net model $\mathcal{N}$ and an observed alarm pattern $\mathcal{A}$, also in Petri net form. We obtain an acyclic net giving partial order representation of the behaviors compatible with the alarm pattern. A recursive online procedure filters out those runs *(configurations)* that explain *exactly*$\mathcal{A}$. The Petri-net based approach generalizes to dynamically evolving topologies, in dynamical systems modeled by graph grammars, see [34]

3.1.2.1. Observability and Diagnosability

Diagnosis algorithms have to operate in contexts with low observability, i.e., in systems where many events are invisible to the supervisor. Checking *observability* and *diagnosability* for the supervised systems is therefore a crucial and non-trivial task in its own right. Analysis of the relational structure of occurrence nets allows us to check whether the system exhibits sufficient visibility to allow diagnosis. Developing efficient methods for both verification of *diagnosability checking* under concurrency, and the *diagnosis* itself for distributed, composite and asynchronous systems, is an important field for *MExICo*. In 2019, a new property, manifestability, weaker than diagnosability (dual in some sense to opacity) has been studied in the context of automata and timed automata.

3.1.2.2. Distribution

Distributed computation of unfoldings allows one to factor the unfolding of the global system into smaller *local* unfoldings, by local supervisors associated with sub-networks and communicating among each other. In [46], [36], elements of a methodology for distributed computation of unfoldings between several supervisors, underwritten by algebraic properties of the category of Petri nets have been developed. Generalizations, in particular to Graph Grammars, are still do be done.

Computing diagnosis in a distributed way is only one aspect of a much vaster topic, that of *distributed diagnosis* (see [43], [49]). In fact, it involves a more abstract and often indirect reasoning to conclude whether or not some given invisible fault has occurred. Combination of local scenarios is in general not sufficient: the global system may have behaviors that do not reveal themselves as faulty (or, dually, non-faulty) on any local supervisor's domain (compare [33], [39]). Rather, the local diagnosers have to join all *information* that is available to them locally, and then deduce collectively further information from the combination of their views. In particular, even the *absence* of fault evidence on all peers may allow to deduce fault occurrence jointly, see [51], [52]. Automatizing such procedures for the supervision and management of distributed and locally monitored asynchronous systems is a long-term goal to which *MExICo* hopes to contribute.

3.1.3. Hybrid Systems

Participants: Philippe Dague, Lina Ye, Serge Haddad.

Hybrid systems constitute a model for cyber-physical systems which integrates continuous-time dynamics (modes) governed by differential equations, and discrete transitions which switch instantaneously from one mode to another. Thanks to their ease of programming, hybrid systems have been integrated to power electronics systems, and more generally in cyber-physical systems. In order to guarantee that such systems meet their specifications, classical methods consist in finitely abstracting the systems by discretization of the (infinite) state space, and deriving automatically the appropriate mode control from the specification using standard graph techniques.

Diagnosability of hybrid systems has also been studied through an abstraction / refinement process in terms of timed automata.

3.1.4. Contextual Nets

Participant: Stefan Schwoon.

Assuring the correctness of concurrent systems is notoriously difficult due to the many unforeseeable ways in which the components may interact and the resulting state-space explosion. A well-established approach to alleviate this problem is to model concurrent systems as Petri nets and analyse their unfoldings, essentially an acyclic version of the Petri net whose simpler structure permits easier analysis [44].

However, Petri nets are inadequate to model concurrent read accesses to the same resource. Such situations often arise naturally, for instance in concurrent databases or in asynchronous circuits. The encoding tricks typically used to model these cases in Petri nets make the unfolding technique inefficient. Contextual nets, which explicitly do model concurrent read accesses, address this problem. Their accurate representation of concurrency makes contextual unfoldings up to exponentially smaller in certain situations. An abstract algorithm for contextual unfoldings was first given in [35]. In recent work, we further studied this subject

from a theoretical and practical perspective, allowing us to develop concrete, efficient data structures and algorithms and a tool (Cunf) that improves upon existing state of the art. This work led to the PhD thesis of César Rodríguez in 2014 .

Contextual unfoldings deal well with two sources of state-space explosion: concurrency and shared resources. Recently, we proposed an improved data structure, called *contextual merged processes* (CMP) to deal with a third source of state-space explosion, i.e. sequences of choices. The work on CMP [53] is currently at an abstract level. In the short term, we want to put this work into practice, requiring some theoretical groundwork, as well as programming and experimentation.

Another well-known approach to verifying concurrent systems is *partial-order reduction*, exemplified by the tool SPIN. Although it is known that both partial-order reduction and unfoldings have their respective strengths and weaknesses, we are not aware of any conclusive comparison between the two techniques. Spin comes with a high-level modeling language having an explicit notion of processes, communication channels, and variables. Indeed, the reduction techniques implemented in Spin exploit the specific properties of these features. On the other side, while there exist highly efficient tools for unfoldings, Petri nets are a relatively general low-level formalism, so these techniques do not exploit properties of higher language features. Our work on contextual unfoldings and CMPs represents a first step to make unfoldings exploit richer models. In the long run, we wish raise the unfolding technique to a suitable high-level modelling language and develop appropriate tool support.

3.2. Management of Quantitative Behavior

Participants: Thomas Chatain, Stefan Haar, Serge Haddad.

3.2.1. Introduction

Besides the logical functionalities of programs, the *quantitative* aspects of component behavior and interaction play an increasingly important role.

- *Real-time* properties cannot be neglected even if time is not an explicit functional issue, since transmission delays, parallelism, etc, can lead to time-outs striking, and thus change even the logical course of processes. Again, this phenomenon arises in telecommunications and web services, but also in transport systems.
- In the same contexts, *probabilities* need to be taken into account, for many diverse reasons such as unpredictable functionalities, or because the outcome of a computation may be governed by race conditions.
- Last but not least, constraints on *cost* cannot be ignored, be it in terms of money or any other limited resource, such as memory space or available CPU time.

Traditional mainframe systems were proprietary and (essentially) localized; therefore, impact of delays, unforeseen failures, etc. could be considered under the control of the system manager. It was therefore natural, in verification and control of systems, to focus on *functional* behavior entirely.

With the increase in size of computing system and the growing degree of compositionality and distribution, quantitative factors enter the stage:

- calling remote services and transmitting data over the web creates *delays*;
- remote or non-proprietary components are not “deterministic”, in the sense that their behavior is uncertain.

Time and *probability* are thus parameters that management of distributed systems must be able to handle; along with both, the *cost* of operations is often subject to restrictions, or its minimization is at least desired. The mathematical treatment of these features in distributed systems is an important challenge, which *MExICo* is addressing; the following describes our activities concerning probabilistic and timed systems. Note that cost optimization is not a current activity but enters the picture in several intended activities.

3.2.2. Probabilistic distributed Systems

Participants: Stefan Haar, Serge Haddad.

3.2.2.1. Non-sequential probabilistic processes

Practical fault diagnosis requires to select explanations of *maximal likelihood*. For partial-order based diagnosis, this leads therefore to the question what the probability of a given partially ordered execution is. In Benveniste et al. [38], [31], we presented a model of stochastic processes, whose trajectories are partially ordered, based on local branching in Petri net unfoldings; an alternative and complementary model based on Markov fields is developed in [48], which takes a different view on the semantics and overcomes the first model's restrictions on applicability.

Both approaches abstract away from real time progress and randomize choices in *logical* time. On the other hand, the relative speed - and thus, indirectly, the real-time behavior of the system's local processes - are crucial factors determining the outcome of probabilistic choices, even if non-determinism is absent from the system.

In another line of research [40] we have studied the likelihood of occurrence of non-sequential runs under random durations in a stochastic Petri net setting. It remains to better understand the properties of the probability measures thus obtained, to relate them with the models in logical time, and exploit them e.g. in *diagnosis*.

3.2.2.2. Distributed Markov Decision Processes

Participant: Serge Haddad.

Distributed systems featuring non-deterministic and probabilistic aspects are usually hard to analyze and, more specifically, to optimize. Furthermore, high complexity theoretical lower bounds have been established for models like partially observed Markovian decision processes and distributed partially observed Markovian decision processes. We believe that these negative results are consequences of the choice of the models rather than the intrinsic complexity of problems to be solved. Thus we plan to introduce new models in which the associated optimization problems can be solved in a more efficient way. More precisely, we start by studying connection protocols weighted by costs and we look for online and offline strategies for optimizing the mean cost to achieve the protocol. We have been cooperating on this subject with the SUMO team at Inria Rennes; in the joint work [32]; there, we strive to synthesize for a given MDP a control so as to guarantee a specific stationary behavior, rather than - as is usually done - so as to maximize some reward.

3.2.3. Large scale probabilistic systems

Addressing large-scale probabilistic systems requires to face state explosion, due to both the discrete part and the probabilistic part of the model. In order to deal with such systems, different approaches have been proposed:

- Restricting the synchronization between the components as in queuing networks allows to express the steady-state distribution of the model by an analytical formula called a product-form [37].
- Some methods that tackle with the combinatory explosion for discrete-event systems can be generalized to stochastic systems using an appropriate theory. For instance symmetry based methods have been generalized to stochastic systems with the help of aggregation theory [42].
- At last simulation, which works as soon as a stochastic operational semantic is defined, has been adapted to perform statistical model checking. Roughly speaking, it consists to produce a confidence interval for the probability that a random path fulfills a formula of some temporal logic [54] .

We want to contribute to these three axes: (1) we are looking for product-forms related to systems where synchronization are more involved (like in Petri nets [6]); (2) we want to adapt methods for discrete-event systems that require some theoretical developments in the stochastic framework and, (3) we plan to address some important limitations of statistical model checking like the expressiveness of the associated logic and the handling of rare events.

3.2.4. Real time distributed systems

Nowadays, software systems largely depend on complex timing constraints and usually consist of many interacting local components. Among them, railway crossings, traffic control units, mobile phones, computer servers, and many more safety-critical systems are subject to particular quality standards. It is therefore becoming increasingly important to look at networks of timed systems, which allow real-time systems to operate in a distributed manner.

Timed automata are a well-studied formalism to describe reactive systems that come with timing constraints. For modeling distributed real-time systems, networks of timed automata have been considered, where the local clocks of the processes usually evolve at the same rate [50] [41]. It is, however, not always adequate to assume that distributed components of a system obey a global time. Actually, there is generally no reason to assume that different timed systems in the networks refer to the same time or evolve at the same rate. Any component is rather determined by local influences such as temperature and workload.

3.2.4.1. Implementation of Real-Time Concurrent Systems

Participants: Thomas Chatain, Stefan Haar, Serge Haddad.

This was one of the tasks of the ANR ImpRo.

Formal models for real-time systems, like timed automata and time Petri nets, have been extensively studied and have proved their interest for the verification of real-time systems. On the other hand, the question of using these models as specifications for designing real-time systems raises some difficulties. One of those comes from the fact that the real-time constraints introduce some artifacts and because of them some syntactically correct models have a formal semantics that is clearly unrealistic. One famous situation is the case of Zeno executions, where the formal semantics allows the system to do infinitely many actions in finite time. But there are other problems, and some of them are related to the distributed nature of the system. These are the ones we address here.

One approach to implementability problems is to formalize either syntactical or behavioral requirements about what should be considered as a reasonable model, and reject other models. Another approach is to adapt the formal semantics such that only realistic behaviors are considered.

These techniques are preliminaries for dealing with the problem of implementability of models. Indeed implementing a model may be possible at the cost of some transformation, which make it suitable for the target device. By the way these transformations may be of interest for the designer who can now use high-level features in a model of a system or protocol, and rely on the transformation to make it implementable.

We aim at formalizing and automating translations that preserve both the timed semantics and the concurrent semantics. This effort is crucial for extending concurrency-oriented methods for logical time, in particular for exploiting partial order properties. In fact, validation and management - in a broad sense - of distributed systems is not realistic *in general* without understanding and control of their real-time dependent features; the link between real-time and logical-time behaviors is thus crucial for many aspects of *MExICo*'s work.

MOCQUA Team

3. Research Program

3.1. Quantum Computing

While it can be argued that the quantum revolution has already happened in cryptography [39] or in optics [38], quantum computers are far from becoming a common commodity, with only a few teams around the world working on a practical implementation. In fact, one of the most commonly known examples of a quantum computer, the D-Wave 2X System, defies the usual definition of a computer: it is not general-purpose, and can only solve (approximately) a very specific hardwired problem.

Most current prototypes of a quantum computer differ fundamentally on the hardware substrate, and it is quite hard to predict which solution will finally be adopted. The landscape of quantum programming languages is also constantly evolving. Comparably to compiler design, the foundation of quantum software therefore relies on an intermediate representation that is suitable for manipulation, easy to produce from software and easily encodable into hardware. The language of choice for this is the ZX-calculus.

Regardless of the actual model that will be accepted by the industry, it is becoming clear that some of the hurdles into scaling up quantum computers from a few qubits to very large arrays will remain. As an example, current implementations of quantum computers working on hundreds of qubits indeed are not able to form and maintain all possible forms of entanglement between qubits. This raises two questions. First, does this restrict the computational power, and the supposed advantage of the quantum computer over the classical computer? Second, how to ensure that a quantum program that was designed for a theoretical quantum computer will work on the practical implementations? This will be investigated, in particular by providing static analysis methods for evaluating a priori how much entanglement a quantum program needs.

3.2. Higher-Order Computing

While programs often operate on natural numbers or finite structures such as graphs or finite strings, they can also take functions as input. In that case, the program is said to perform higher-order computations, or to compute a higher-order functional. Functional programming or object-oriented programming are important paradigms allowing higher-order computations.

While the theory of computation is well developed for first-order programs, difficulties arise when dealing with higher-order programs. There are many non-equivalent ways of presenting inputs to such programs: an input function can be presented as a black-box, encoded in an infinite binary sequence, or sometimes by a finite description. Comparing those representations is an important problem. A particularly useful application of higher-order computations is to compute with infinite objects that can be represented by functions or symbolic sequences. The theory works well in many cases (to be precise, when these objects live in a topological space with a countable basis [42]), but is not well understood in other interesting cases. For instance, when the inputs are the second-order functionals (of type $(\mathbb{N} \to \mathbb{N}) \to (\mathbb{N} \to \mathbb{N})$), the classical theory does not apply and many problems are still open.

3.3. Dynamical Systems

The most natural example of a computation with infinite precision is the simulation of a dynamical system. The underlying space might be $\mathbb{R}^n$ in the case of the simulation of physical systems, or the Cantor space $\{0,1\}^{\mathbb{Z}}$ in the case of discrete dynamical systems.

From the point of view of computation, the main point of interest is the link between the long-term behavior of a system and its initial configuration. There are two questions here: (a) predict the behavior, (b) design dynamical systems with some prescribed behavior. The first will be mainly examined through the angle of reachability and more generally control theory for hybrid systems.

The model of cellular automata will be of particular interest. This computational model is relevant for simulating complex global phenomena which emerge from simple interactions between simple components. It is widely used in various natural sciences (physics, biology, etc.) and in computer science, as it is an appropriate model to reason about errors that occur in systems with a great number of components.

The simulation of a physical dynamical system on a computer is made difficult by various aspects. First, the parameters of the dynamical systems are seldom exactly known. Secondly, the simulation is usually non exact: real numbers are usually represented by floating-point numbers, and simulations of cellular automata only simulate the behavior of finite or periodic configurations. For some chaotic systems, this means that the simulation can be completely irrelevant.

OURAGAN Project-Team

3. Research Program

3.1. Basic computable objects and algorithms

The development of basic computable objects is somehow *on demand* and depends on all the other directions. However, some critical computations are already known to be bottlenecks and are sources of constant efforts.

Computations with algebraic numbers appear in almost all our activities: when working with number fields in our work in algorithmic number theory as well as in all the computations that involve the use of solutions of zero-dimensional systems of polynomial equations. Among the identified problems: finding good representations for single number fields (optimizing the size and degree of the defining polynomials), finding good representations for towers or products of number fields (typically working with a tower or finding a unique good extension), efficiently computing in practice with number fields (using certified approximation vs working with the formal description based on polynomial arithmetics). Strong efforts are currently done in the understanding of the various strategies by means of tight theoretical complexity studies [70], [115], [50] and many other efforts will be required to find the right representation for the right problem in practice. For example, for isolating critical points of plane algebraic curves, it is still unclear (at least the theoretical complexity cannot help) that an intermediate formal parameterization is more efficient than a triangular decomposition of the system and it is still unclear that these intermediate computations could be dominated in time by the certified final approximation of the roots.

3.2. Algorithmic Number Theory

Concerning algorithmic number theory, the main problems we will be considering in the coming years are the following:

- *Number fields.* We will continue working on the problems of class groups and generators. In particular, the existence and accessibility of *good* defining polynomials for a fixed number field remain very largely open. The impact of better polynomials on the algorithmic performance is a very important parameter, which makes this problem essential.
- *Lattice reduction.* Despite a great amount of work in the past 35 years on the LLL algorithm and its successors, many open problems remain. We will continue the study of the use of interval arithmetic in this field and the analysis of variants of LLL along the lines of the *Potential*-LLL which provides improved reduction comparable to BKZ with a small block size but has better performance.
- *Elliptic curves and Drinfeld modules.* The study of elliptic curves is a very fruitful area of number theory with many applications in crypto and algorithms. Drinfeld modules are "cousins" of elliptic curves which have been less explored in the algorithm context. However, some recent advances [74] have used them to provide some fast sophisticated factoring algorithms. As a consequence, it is natural to include these objects in our research directions.

3.2.1. Rigorous numerical computations

Some studies in this area will be driven by some other directions, for example, the rigorous evaluation of non algebraic functions on algebraic varieties might become central for some of our work on topology in small dimension (volumes of varieties, drawing of amoeba) or control theory (approximations of discriminant varieties) are our two main current sources of interesting problems. In the same spirit, the work on L-functions computations (extending the computation range, algorithmic tools for computing algebraic data from the L function) will naturally follow.

On the other hand, another objective is to extend existing results on periods of algebraic curves to general curves and higher dimensional varieties is a general promising direction. This project aims at providing tools for integration on higher homology groups of algebraic curves, ie computing Gauss-Manin connections. It requires good understanding of their topology, and more algorithmic tools on differential equations.

3.3. Topology in small dimension

3.3.1. Character varieties

The brute force approach to computable objects from topology of small dimension will not allow any significant progress. As explained above, the systems that arise from these problems are simply outside the range of doable computations. We still continue the work in this direction by a four-fold approach, with all three directions deeply inter-related. First, we focus on a couple of especially meaningful (for the applications) cases, in particular the 3-dimensional manifold called Whitehead link complement. At this point, we are able to make steps in the computation and describe part of the solutions [79], [89]; we hope to be able to complete the computation using every piece of information to simplify the system. Second, we continue the theoretical work to understand more properties of these systems [77]. These properties may prove how useful for the mathematical understanding is the resolution of such systems - or at least the extraction of meaningful information. This approach is for example carried on by Falbel and his work on configuration of flags [80], [82]. Third, we position ourselves as experts in the know-how of this kind of computations and natural interlocutors for colleagues coming up with a question on such a computable object (see [87] and [89]). This also allows us to push forward the kind of computation we actually do and make progress in the direction of the second point. We are credible interlocutors because our team has the blend of theoretical knowledge and computational capabilities that grants effective resolutions of the problems we are presented. And last, we use the knowledge already acquired to pursue our theoretical study of the CR-spherical geometry [69], [81], [78].

Another direction of work is the help to the community in experimental mathematics on new objects. It involves downsizing the system we are looking at (for example by going back to systems coming from hyperbolic geometry and not CR-spherical geometry) and get the most out of what we can compute, by studying new objects. An example of this research direction is the work of Guilloux around the volume function on deformation varieties. This is a real-analytic function defined on the varieties we specialized in computing. Being able to do effective computations with this function led first to a conjecture [86]. Then, theoretical discussions around this conjecture led to a paper on a new approach to the Mahler measure of some 2-variables polynomials [88]. In turn, this last paper gave a formula for the Mahler measure in terms of a function akin to the volume function applied at points in an algebraic variety whose moduli of coordinates are 1. The OURAGAN team has the expertise to compute all the objects appearing in this formula, opening the way to another area of application. This area is deeply linked with number theory as well as topology of small dimension. It requires all the tools at disposition within OURAGAN.

3.3.2. Knot theory

We will carry on the exhaustive search for the lexicographic degrees for the rational knots. They correspond to trigonal space curves: computations in the braid group B_3, explicit parametrization of trigonal curves corresponding to "dessins d'enfants", etc. The problem seems much more harder when looking for more general knots.

On the other hand, a natural direction would be: given an explicit polynomial space curve, determine the under/over nature of the crossings when projecting, draw it and determine the known knot [0] it is isotopic to.

3.3.3. Vizualisation and Computational Geometry

As mentioned above, the drawing of algebraic curves and surfaces is a critical action in OURAGAN since it is a key ingredient in numerous developments. In some cases, one will need a fully certified study of the variety for deciding existence of solutions (for example a region in a robot's parameter's space with solutions

[0] for example the first rational knots are listed at https://team.Inria.fr/ouragan/knots

to the DKP above or deciding if some variety crosses the unit polydisk for some stability problems in control-theory), in some other cases just a partial but certified approximation of a surface (path planning in robotics, evaluation of non algebraic functions over an algebraic variety for volumes of knot complements in the study of character varieties).

On the one hand, we will contribute to general tools like ISOTOP [0] under the supervision of the GAMBLE project-team and, on the other hand, we will propose ad-hoc solutions by gluing some of our basic tools (problems of high degrees in robust control theory). The priority is to provide a first software that implements methods that fit as most as possible the very last complexity results we got on several (theoretical) algorithms for the computation of the topology of plane curves.

A particular effort will be devoted to the resolution of overconstraint bivariate systems which are useful for the studies of singular points and to polynomials systems in 3 variables in the same spirit : avoid the use of Gröbner basis and propose a new algorithm with a state-of-the-art complexity and with a good practical behavior.

In parallel, one will have to carefully study the drawing of graphs of non algebraic functions over algebraic complex surfaces for providing several tools which are useful for mathematicians working on topology in small dimension (a well known example is the drawing of amoebia, a way of representing a complex curve on a sheet of paper).

3.4. Algebraic analysis of functional systems

We want to further develop our expertise in the computational aspects of algebraic analysis by continuing to develop effective versions of results of module theory, homological algebra, category theory and sheaf theory [136] which play important roles in algebraic analysis [45], [103], [104] and in the algorithmic study of linear functional systems. In particular, we shall focus on linear systems of integro-differential-constant/varying/distributed delay equations [124], [126] which play an important role in mathematical systems theory, control theory, and signal processing [124], [131], [125], [128].

The rings of integro-differential operators are highly more complicated than the purely differential case (i.e. Weyl algebras) [12], due to the existence of zero-divisors, or the fact of having a coherent ring instead of a noetherian ring [42]. Therefore, we want to develop an algorithmic study of these rings. Following the direction initiated in [126] for the computation of zero divisors (based on the polynomial null spaces of certain operators), we first want to develop algorithms for the computation of left/right kernels and left/right/generalized inverses of matrices with entries in such rings, and to use these results in module theory (e.g. computation of syzygy modules, (shorter/shortest) free resolutions, split short/long exact sequences). Moreover, Stafford's results [137], algorithmically developed in [12] for rings of partial differential operators (i.e. the Weyl algebras), are known to still hold for rings of integro-differential operators. We shall study their algorithmic extensions. Our corresponding implementation will be extended accordingly.

Finally, within a computer algebra viewpoint, we shall continue to algorithmically study issues on rings of integro-differential-delay operators [124], [125] and their applications to the study of equivalences of differential constant/varying/distributed delay systems (e.g. Artstein's reduction, Fiagbedzi-Pearson's transformation) which play an important role in control theory.

[0] https://isotop.gamble.loria.fr

PACAP Project-Team

3. Research Program

3.1. Motivation

Our research program is naturally driven by the evolution of our ecosystem. Relevant recent changes can be classified in the following categories: technological constraints, evolving community, and domain constraints. We hereby summarize these evolutions.

3.1.1. Technological constraints

Until recently, binary compatibility guaranteed portability of programs, while increased clock frequency and improved micro-architecture provided increased performance. However, in the last decade, advances in technology and micro-architecture started translating into more parallelism instead. Technology roadmaps even predict the feasibility of thousands of cores on a chip by 2020. Hundreds are already commercially available. Since the vast majority of applications are still sequential, or contain significant sequential sections, such a trend put an end to the automatic performance improvement enjoyed by developers and users. Many research groups consequently focused on parallel architectures and compiling for parallelism.

Still, the performance of applications will ultimately be driven by the performance of the sequential part. Despite a number of advances (some of them contributed by members of the team), sequential tasks are still a major performance bottleneck. Addressing it is still on the agenda of the PACAP project-team.

In addition, due to power constraints, only part of the billions of transistors of a microprocessor can be operated at any given time (the *dark silicon* paradigm). A sensible approach consists in specializing parts of the silicon area to provide dedicated accelerators (not run simultaneously). This results in diverse and heterogeneous processor cores. Application and compiler designers are thus confronted with a moving target, challenging portability and jeopardizing performance.

Note on technology.
Technology also progresses at a fast pace. We do not propose to pursue any research on technology *per se*. Recently proposed paradigms (non-Silicon, brain-inspired) have received lots of attention from the research community. We do *not* intend to invest in those paradigms, but we will continue to investigate compilation and architecture for more conventional programming paradigms. Still, several technological shifts may have consequences for us, and we will closely monitor their developments. They include for example non-volatile memory (impacts security, makes writes longer than loads), 3D-stacking (impacts bandwidth), and photonics (impacts latencies and connection network), quantum computing (impacts the entire software stack).

3.1.2. Evolving community

The PACAP project-team tackles performance-related issues, for conventional programming paradigms. In fact, programming complex environments is no longer the exclusive domain of experts in compilation and architecture. A large community now develops applications for a wide range of targets, including mobile "apps", cloud, multicore or heterogeneous processors.

This also includes domain scientists (in biology, medicine, but also social sciences) who started relying heavily on computational resources, gathering huge amounts of data, and requiring a considerable amount of processing to analyze them. Our research is motivated by the growing discrepancy between on the one hand, the complexity of the workloads and the computing systems, and on the other hand, the expanding community of developers at large, with limited expertise to optimize and to map efficiently computations to compute nodes.

3.1.3. Domain constraints

Mobile, embedded systems have become ubiquitous. Many of them have real-time constraints. For this class of systems, correctness implies not only producing the correct result, but also doing so within specified deadlines. In the presence of heterogeneous, complex and highly dynamic systems, producing *tight* (i.e., useful) upper bound to the worst-case execution time has become extremely challenging. Our research will aim at improving the tightness as well as enlarging the set of features that can be safely analyzed.

The ever growing dependence of our economy on computing systems also implies that security has become of utmost importance. Many systems are under constant attacks from intruders. Protection has a cost also in terms of performance. We plan to leverage our background to contribute solutions that minimize this impact.

Note on Applications Domains.
PACAP works on fundamental technologies for computer science: processor architecture, performance-oriented compilation and guaranteed response time for real-time. The research results may have impact on any application domain that requires high performance execution (telecommunication, multimedia, biology, health, engineering, environment...), but also on many embedded applications that exhibit other constraints such as power consumption, code size and guaranteed response time.

We strive to extract from active domains the fundamental characteristics that are relevant to our research. For example, *big data* is of interest to PACAP because it relates to the study of hardware/software mechanisms to efficiently transfer huge amounts of data to the computing nodes. Similarly, the *Internet of Things* is of interest because it has implications in terms of ultra low-power consumption.

3.2. Research Objectives

Processor micro-architecture and compilation have been at the core of the research carried by the members of the project teams for two decades, with undeniable contributions. They continue to be the foundation of PACAP.

Heterogeneity and diversity of processor architectures now require new techniques to guarantee that the hardware is satisfactorily exploited by the software. One of our goals is to devise new static compilation techniques (cf. Section 3.2.1), but also build upon iterative [1] and split [40] compilation to continuously adapt software to its environment (Section 3.2.2). Dynamic binary optimization will also play a key role in delivering adapting software and increased performance.

The end of Moore's law and Dennard's scaling [0] offer an exciting window of opportunity, where performance improvements will no longer derive from additional transistor budget or increased clock frequency, but rather come from breakthroughs in micro-architecture (Section 3.2.3). Reconciling CPU and GPU designs (Section 3.2.4) is one of our objectives.

Heterogeneity and multicores are also major obstacles to determining tight worst-case execution times of real-time systems (Section 3.2.5), which we plan to tackle.

Finally, we also describe how we plan to address transversal aspects such as power efficiency (Section 3.2.6), and security (Section 3.2.7).

3.2.1. Static Compilation

Static compilation techniques continue to be relevant in addressing the characteristics of emerging hardware technologies, such as non-volatile memories, 3D-stacking, or novel communication technologies. These techniques expose new characteristics to the software layers. As an example, non-volatile memories typically have asymmetric read-write latencies (writes are much longer than reads) and different power consumption profiles. PACAP studies new optimization opportunities and develops tailored compilation techniques for upcoming compute nodes. New technologies may also be coupled with traditional solutions to offer new

[0] According to Dennard scaling, as transistors get smaller the power density remains constant, and the consumed power remains proportional to the area.

trade-offs. We study how programs can adequately exploit the specific features of the proposed heterogeneous compute nodes.

We propose to build upon iterative compilation [1] to explore how applications perform on different configurations. When possible, Pareto points are related to application characteristics. The best configuration, however, may actually depend on runtime information, such as input data, dynamic events, or properties that are available only at runtime. Unfortunately a runtime system has little time and means to determine the best configuration. For these reasons, we also leverage split-compilation [40]: the idea consists in pre-computing alternatives, and embedding in the program enough information to assist and drive a runtime system towards to the best solution.

3.2.2. Software Adaptation

More than ever, software needs to adapt to its environment. In most cases, this environment remains unknown until runtime. This is already the case when one deploys an application to a cloud, or an "app" to mobile devices. The dilemma is the following: for maximum portability, developers should target the most general device; but for performance they would like to exploit the most recent and advanced hardware features. JIT compilers can handle the situation to some extent, but binary deployment requires dynamic binary rewriting. Our work has shown how SIMD instructions can be upgraded from SSE to AVX transparently [2]. Many more opportunities will appear with diverse and heterogeneous processors, featuring various kinds of accelerators.

On shared hardware, the environment is also defined by other applications competing for the same computational resources. It becomes increasingly important to adapt to changing runtime conditions, such as the contention of the cache memories, available bandwidth, or hardware faults. Fortunately, optimizing at runtime is also an opportunity, because this is the first time the program is visible as a whole: executable and libraries (including library versions). Optimizers may also rely on dynamic information, such as actual input data, parameter values, etc. We have already developed a software platform [46] to analyze and optimize programs at runtime, and we started working on automatic dynamic parallelization of sequential code, and dynamic specialization.

We started addressing some of these challenges in ongoing projects such as Nano2017 PSAIC Collaborative research program with STMicroelectronics, as well as within the Inria Project Lab MULTICORE. The H2020 FET HPC project ANTAREX also addresses these challenges from the energy perspective. We further leverage our platform and initial results to address other adaptation opportunities. Efficient software adaptation requires expertise from all domains tackled by PACAP, and strong interaction between all team members is expected.

3.2.3. Research directions in uniprocessor micro-architecture

Achieving high single-thread performance remains a major challenge even in the multicore era (Amdahl's law). The members of the PACAP project-team have been conducting research in uniprocessor micro-architecture research for about 20 years covering major topics including caches, instruction front-end, branch prediction, out-of-order core pipeline, and value prediction. In particular, in recent years they have been recognized as world leaders in branch prediction [50] [44] and in cache prefetching [6] and they have revived the forgotten concept of value prediction [9][8]. This research was supported by the ERC Advanced grant DAL (2011-2016) and also by Intel. We pursue research on achieving ultimate unicore performance. Below are several non-orthogonal directions that we have identified for mid-term research:

1. management of the memory hierarchy (particularly the hardware prefetching);
2. practical design of very wide issue execution cores;
3. speculative execution.

Memory design issues:
Performance of many applications is highly impacted by the memory hierarchy behavior. The interactions between the different components in the memory hierarchy and the out-of-order execution engine have high impact on performance.

The last *Data Prefetching Contest* held with ISCA 2015 has illustrated that achieving high prefetching efficiency is still a challenge for wide-issue superscalar processors, particularly those featuring a very large instruction window. The large instruction window enables an implicit data prefetcher. The interaction between this implicit hardware prefetcher and the explicit hardware prefetcher is still relatively mysterious as illustrated by Pierre Michaud's BO prefetcher (winner of DPC2) [6]. The first research objective is to better understand how the implicit prefetching enabled by the large instruction window interacts with the L2 prefetcher and then to understand how explicit prefetching on the L1 also interacts with the L2 prefetcher.

The second research objective is related to the interaction of prefetching and virtual/physical memory. On real hardware, prefetching is stopped by page frontiers. The interaction between TLB prefetching (and on which level) and cache prefetching must be analyzed.

The prefetcher is not the only actor in the hierarchy that must be carefully controlled. Significant benefits can also be achieved through careful management of memory access bandwidth, particularly the management of spatial locality on memory accesses, both for reads and writes. The exploitation of this locality is traditionally handled in the memory controller. However, it could be better handled if larger temporal granularity was available. Finally, we also intend to continue to explore the promising avenue of compressed caches. In particular we recently proposed the skewed compressed cache [11]. It offers new possibilities for efficient compression schemes.

Ultra wide-issue superscalar.
To effectively leverage memory level parallelism, one requires huge out-of-order execution structures as well as very wide issue superscalar processors. For the two past decades, implementing ever wider issue superscalar processors has been challenging. The objective of our research on the execution core is to explore (and revisit) directions that allow the design of a very wide-issue (8-to-16 way) out-of-order execution core while mastering its complexity (silicon area, hardware logic complexity, power/energy consumption).

The first direction that we are exploring is the use of clustered architectures [7]. Symmetric clustered organization allows to benefit from a simpler bypass network, but induce large complexity on the issue queue. One remarkable finding of our study [7] is that, when considering two large clusters (e.g. 8-wide), steering large groups of consecutive instructions (e.g. 64 μops) to the same cluster is quite efficient. This opens opportunities to limit the complexity of the issue queues (monitoring fewer buses) and register files (fewer ports and physical registers) in the clusters, since not all results have to be forwarded to the other cluster.

The second direction that we are exploring is associated with the approach that we developed with Sembrant et al. [47]. It reduces the number of instructions waiting in the instruction queues for the applications benefiting from very large instruction windows. Instructions are dynamically classified as ready (independent from any long latency instruction) or non-ready, and as urgent (part of a dependency chain leading to a long latency instruction) or non-urgent. Non-ready non-urgent instructions can be delayed until the long latency instruction has been executed; this allows to reduce the pressure on the issue queue. This proposition opens the opportunity to consider an asymmetric micro-architecture with a cluster dedicated to the execution of urgent instructions and a second cluster executing the non-urgent instructions. The micro-architecture of this second cluster could be optimized to reduce complexity and power consumption (smaller instruction queue, less aggressive scheduling...)

Speculative execution.
Out-of-order (OoO) execution relies on speculative execution that requires predictions of all sorts: branch, memory dependency, value...

The PACAP members have been major actors of branch prediction research for the last 20 years; and their proposals have influenced the design of most of the hardware branch predictors in current microprocessors. We will continue to steadily explore new branch predictor designs, as for instance [48].

In speculative execution, we have recently revisited value prediction (VP) which was a hot research topic between 1996 and 2002. However it was considered until recently that value prediction would lead to a huge increase in complexity and power consumption in every stage of the pipeline. Fortunately, we have recently shown that complexity usually introduced by value prediction in the OoO engine can be overcome [9][8] [50] [44]. First, very high accuracy can be enforced at reasonable cost in coverage and minimal complexity [9]. Thus, both prediction validation and recovery by squashing can be done outside the out-of-order engine, at commit time. Furthermore, we propose a new pipeline organization, EOLE ({Early | Out-of-order | Late} Execution), that leverages VP with validation at commit to execute many instructions outside the OoO core, in-order [8]. With EOLE, the issue-width in OoO core can be reduced without sacrificing performance, thus benefiting the performance of VP without a significant cost in silicon area and/or energy. In the near future, we will explore new avenues related to value prediction. These directions include register equality prediction and compatibility of value prediction with weak memory models in multiprocessors.

3.2.4. Towards heterogeneous single-ISA CPU-GPU architectures

Heterogeneous single-ISA architectures have been proposed in the literature during the 2000's [43] and are now widely used in the industry (Arm big.LITTLE, NVIDIA 4+1...) as a way to improve power-efficiency in mobile processors. These architectures include multiple cores whose respective micro-architectures offer different trade-offs between performance and energy efficiency, or between latency and throughput, while offering the same interface to software. Dynamic task migration policies leverage the heterogeneity of the platform by using the most suitable core for each application, or even each phase of processing. However, these works only tune cores by changing their complexity. Energy-optimized cores are either identical cores implemented in a low-power process technology, or simplified in-order superscalar cores, which are far from state-of-the-art throughput-oriented architectures such as GPUs.

We investigate the convergence of CPU and GPU at both architecture and compiler levels.

Architecture.
The architecture convergence between Single Instruction Multiple Threads (SIMT) GPUs and multicore processors that we have been pursuing [42] opens the way for heterogeneous architectures including latency-optimized superscalar cores and throughput-optimized GPU-style cores, which all share the same instruction set. Using SIMT cores in place of superscalar cores will enable the highest energy efficiency on regular sections of applications. As with existing single-ISA heterogeneous architectures, task migration will not necessitate any software rewrite and will accelerate existing applications.

Compilers for emerging heterogeneous architectures.
Single-ISA CPU+GPU architectures will provide the necessary substrate to enable efficient heterogeneous processing. However, it will also introduce substantial challenges at the software and firmware level. Task placement and migration will require advanced policies that leverage both static information at compile time and dynamic information at run-time. We are tackling the heterogeneous task scheduling problem at the compiler level.

3.2.5. Real-time systems

Safety-critical systems (e.g. avionics, medical devices, automotive...) have so far used simple unicore hardware systems as a way to control their predictability, in order to meet timing constraints. Still, many critical embedded systems have increasing demand in computing power, and simple unicore processors are not sufficient anymore. General-purpose multicore processors are not suitable for safety-critical real-time systems, because they include complex micro-architectural elements (cache hierarchies, branch, stride and value predictors) meant to improve average-case performance, and for which worst-case performance is difficult to predict. The prerequisite for calculating tight WCET is a deterministic hardware system that avoids dynamic, time-unpredictable calculations at run-time.

Even for multi and manycore systems designed with time-predictability in mind (Kalray MPPA manycore architecture [0], or the Recore manycore hardware [0]) calculating WCETs is still challenging. The following two challenges will be addressed in the mid-term:

1. definition of methods to estimate WCETs tightly on manycores, that smartly analyze and/or control shared resources such as buses, NoCs or caches;
2. methods to improve the programmability of real-time applications through automatic parallelization and optimizations from model-based designs.

3.2.6. Power efficiency

PACAP addresses power-efficiency at several levels. First, we design static and split compilation techniques to contribute to the race for Exascale computing (the general goal is to reach 10^{18} FLOP/s at less than 20 MW). Second, we focus on high-performance low-power embedded compute nodes. Within the ANR project Continuum, in collaboration with architecture and technology experts from LIRMM and the SME Cortus, we research new static and dynamic compilation techniques that fully exploit emerging memory and NoC technologies. Finally, in collaboration with the CAIRN project-team, we investigate the synergy of reconfigurable computing and dynamic code generation.

Green and heterogeneous high-performance computing.
Concerning HPC systems, our approach consists in mapping, runtime managing and autotuning applications for green and heterogeneous High-Performance Computing systems up to the Exascale level. One key innovation of the proposed approach consists of introducing a separation of concerns (where self-adaptivity and energy efficient strategies are specified aside to application functionalities) promoted by the definition of a Domain Specific Language (DSL) inspired by aspect-oriented programming concepts for heterogeneous systems. The new DSL will be introduced for expressing adaptivity/energy/performance strategies and to enforce at runtime application autotuning and resource and power management. The goal is to support the parallelism, scalability and adaptability of a dynamic workload by exploiting the full system capabilities (including energy management) for emerging large-scale and extreme-scale systems, while reducing the Total Cost of Ownership (TCO) for companies and public organizations.

High-performance low-power embedded compute nodes.
We will address the design of next generation energy-efficient high-performance embedded compute nodes. It focuses at the same time on software, architecture and emerging memory and communication technologies in order to synergistically exploit their corresponding features. The approach of the project is organized around three complementary topics: 1) compilation techniques; 2) multicore architectures; 3) emerging memory and communication technologies. PACAP will focus on the compilation aspects, taking as input the software-visible characteristics of the proposed emerging technology, and making the best possible use of the new features (non-volatility, density, endurance, low-power).

Hardware Accelerated JIT Compilation.
Reconfigurable hardware offers the opportunity to limit power consumption by dynamically adjusting the number of available resources to the requirements of the running software. In particular, VLIW processors can adjust the number of available issue lanes. Unfortunately, changing the processor width often requires recompiling the application, and VLIW processors are highly dependent of the quality of the compilation, mainly because of the instruction scheduling phase performed by the compiler. Another challenge lies in the high constraints of the embedded system: the energy and execution time overhead due to the JIT compilation must be carefully kept under control.

We started exploring ways to reduce the cost of JIT compilation targeting VLIW-based heterogeneous many-core systems. Our approach relies on a hardware/software JIT compiler framework. While basic optimizations and JIT management are performed in software, the compilation back-end is implemented by means of specialized hardware. This back-end involves both instruction scheduling and register allocation, which are known to be the most time-consuming stages of such a compiler.

[0] http://www.kalrayinc.com
[0] http://www.recoresystems.com/

3.2.7. Security

Security is a mandatory concern of any modern computing system. Various threat models have led to a multitude of protection solutions. Members of PACAP already contributed in the past, thanks to the HAVEGE [49] random number generator, and code obfuscating techniques (the obfuscating just-in-time compiler [41], or thread-based control flow mangling [45]). Still, security is not core competence of PACAP members.

Our strategy consists in partnering with security experts who can provide intuition, know-how and expertise, in particular in defining threat models, and assessing the quality of the solutions. Our expertise in compilation and architecture helps design more efficient and less expensive protection mechanisms.

Examples of collaborations so far include the following:

Compilation: We partnered with experts in security and codes to prototype a platform that demonstrates resilient software. They designed and proposed advanced masking techniques to hide sensitive data in application memory. PACAP's expertise is key to select and tune the protection mechanisms developed within the project, and to propose safe, yet cost-effective solutions from an implementation point of view.

Dynamic Binary Rewriting: Our expertise in dynamic binary rewriting combines well with the expertise of the CIDRE team in protecting application. Security has a high cost in terms of performance, and static insertion of counter measures cannot take into account the current threat level. In collaboration with CIDRE, we propose an adaptive insertion/removal of countermeasures in a running application based of dynamic assessment of the threat level.

WCET Analysis: Designing real-time systems requires computing an upper bound of the worst-case execution time. Knowledge of this timing information opens an opportunity to detect attacks on the control flow of programs. In collaboration with CIDRE, we are developing a technique to detect such attacks thanks to a hardware monitor that makes sure that statically computed time information is preserved (CAIRN is also involved in the definition of the hardware component).

PARKAS Project-Team

3. Research Program

3.1. Programming Languages for Cyber-Physical Systems

We study the definition of languages for reactive and Cyber-Physical Systems in which distributed control software interacts closely with physical devices. We focus on languages that mix discrete-time and continuous-time; in particular, the combination of synchronous programming constructs with differential equations, relaxed models of synchrony for distributed systems communicating via periodic sampling or through buffers, and the embedding of synchronous features in a general purpose ML language.

The synchronous language SCADE, [0] based on synchronous languages principles, is ideal for programming embedded software and is used routinely in the most critical applications. But embedded design also involves modeling the control software together with its environment made of physical devices that are traditionally defined by differential equations that evolve on a continuous-time basis and approximated with a numerical solver. Furthermore, compilation usually produces single-loop code, but implementations increasingly involve multiple and multi-core processors communicating via buffers and shared-memory.

The major player in embedded design for cyber-physical systems is undoubtedly SIMULINK, [0] with MODELICA[0] a new player. Models created in these tools are used not only for simulation, but also for test-case generation, formal verification, and translation to embedded code. That said, many foundational and practical aspects are not well-treated by existing theory (for instance, hybrid automata), and current tools. In particular, features that mix discrete and continuous time often suffer from inadequacies and bugs. This results in a broken development chain: for the most critical applications, the model of the controller must be reprogrammed into either sequential or synchronous code, and properties verified on the source model have to be reverified on the target code. There is also the question of how much confidence can be placed in the code used for simulation.

We attack these issues through the development of the ZELUS research prototype, industrial collaborations with the SCADE team at ANSYS/Esterel-Technologies, and collaboration with Modelica developers at Dassault-Systèmes and the Modelica association. Our approach is to develop a *conservative extension* of a synchronous language capable of expressing in a single source text a model of the control software and its physical environment, to simulate the whole using off-the-shelf numerical solvers, and to generate target embedded code. Our goal is to increase faithfulness and confidence in both what is actually executed on platforms and what is simulated. The goal of building a language on a strong mathematical basis for hybrid systems is shared with the Ptolemy project at UC Berkeley; our approach is distinguished by building our language on a synchronous semantics, reusing and extending classical synchronous compilation techniques.

Adding continuous time to a synchronous language gives a richer programming model where reactive controllers can be specified in idealized physical time. An example is the so called quasi-periodic architecture studied by Caspi, where independent processors execute periodically and communicate by sampling. We have applied ZELUS to model a class of quasi-periodic protocols and to analyze an abstraction proposed for model-checking such systems.

Communication-by-sampling is suitable for control applications where value timeliness is paramount and lost or duplicate values tolerable, but other applications—for instance, those involving video streams—seek a different trade-off through the use of bounded buffers between processes. We developed the *n*-synchronous model and the programming language LUCY-N to treat this issue.

[0] http://www.esterel-technologies.com/products/scade-suite
[0] http://www.mathworks.com/products/simulink
[0] https://www.modelica.org

3.2. Efficient Compilation for Parallel and Distributed Computing

We develop compilation techniques for sequential and multi-core processors, and efficient parallel run-time systems for computationally intensive real-time applications (e.g., video and streaming). We study the generation of parallel code from synchronous programs, compilation techniques based on the polyhedral model, and the exploitation of synchronous Single Static Assignment (SSA) representations in general purpose compilers.

We consider distribution and parallelism as two distinct concepts.

- Distribution refers to the construction of multiple programs which are dedicated to run on specific computing devices. When an application is designed for, or adapted to, an embedded multiprocessor, the distribution task grants fine grained—design- or compilation-time—control over the mapping and interaction between the multiple programs.
- Parallelism is about generating code capable of efficiently exploiting multiprocessors. Typically this amounts to maing (in)dependence properties, data transfers, atomicity and isolation explicit. Compiling parallelism translates these properties into low-level synchronization and communication primitives and/or onto a runtime system.

We also see a strong relation between the foundations of synchronous languages and the design of compiler intermediate representations for concurrent programs. These representations are essential to the construction of compilers enabling the optimization of parallel programs and the management of massively parallel resources. Polyhedral compilation is one of the most popular research avenues in this area. Indirectly, the design of intermediate representations also triggers exciting research on dedicated runtime systems supporting parallel constructs. We are particularly interested in the implementation of non-blocking dynamic schedulers interacting with decoupled, deterministic communication channels to hide communication latency and optimize local memory usage.

While distribution and parallelism issues arise in all areas of computing, our programming language perspective pushes us to consider four scenarios:

1. designing an embedded system, both hardware and software, and codesign;
2. programming existing embedded hardware with functional and behavioral constraints;
3. programming and compiling for a general-purpose or high-performance, best-effort system;
4. programming large scale distributed, I/O-dominated and data-centric systems.

We work on a multitude of research experiments, algorithms and prototypes related to one or more of these scenarios. Our main efforts focused on extending the code generation algorithms for synchronous languages and on the development of more scalable and widely applicable polyhedral compilation methods.

3.3. Validation and Proof of Compilers

Compilers are complex software and not immune from bugs. We work on validation and proof tools for compilers to relate the semantics of executed code and source programs. We develop techniques to formally prove the correctness of compilation passes for synchronous languages (Lustre), and to validate compilation optimization for C code in the presence of threads.

3.3.1. Lustre:

The formal validation of a compiler for a synchronous language (or more generally for a language based on synchronous block diagrams) promises to reduce the likelihood of compiler-introduced bugs, the cost of testing, and also to ensure that properties verified on the source model hold of the target code. Such a validation would be complementary to existing industrial qualifications which certify the development process and not the functional correctness of a compiler. The scientific interest is in developing models and techniques that both facilitate the verification and allow for convenient reasoning over the semantics of a language and the behavior of programs written in it.

3.3.2. C/C++:

The recently approved C11 and C++11 standards define a concurrency model for the C and C++ languages, which were originally designed without concurrency support. Their intent is to permit most compiler and hardware optimizations, while providing escape mechanisms for writing portable, high-performance, low-level code. Mainstream compilers are being modified to support the new standards. A subtle class of compiler bugs is the so-called concurrency compiler bugs, where compilers generate correct sequential code but break the concurrency memory model of the programming language. Such bugs are observable only when the miscompiled functions interact with concurrent contexts, making them particularly hard to detect. All previous techniques to test compiler correctness miss concurrency compiler bugs.

PARSIFAL Project-Team

3. Research Program

3.1. General overview

There are two broad approaches for computational specifications. In the *computation as model* approach, computations are encoded as mathematical structures containing nodes, transitions, and state. Logic is used to *describe* these structures, that is, the computations are used as models for logical expressions. Intensional operators, such as the modals of temporal and dynamic logics or the triples of Hoare logic, are often employed to express propositions about the change in state.

The *computation as deduction* approach, in contrast, expresses computations logically, using formulas, terms, types, and proofs as computational elements. Unlike the model approach, general logical apparatus such as cut-elimination or automated deduction becomes directly applicable as tools for defining, analyzing, and animating computations. Indeed, we can identify two main aspects of logical specifications that have been very fruitful:

- *Proof normalization*, which treats the state of a computation as a proof term and computation as normalization of the proof terms. General reduction principles such as β-reduction or cut-elimination are merely particular forms of proof normalization. Functional programming is based on normalization [51], and normalization in different logics can justify the design of new and different functional programming languages [32].
- *Proof search*, which views the state of a computation as a a structured collection of formulas, known as a *sequent*, and proof search in a suitable sequent calculus as encoding the dynamics of the computation. Logic programming is based on proof search [55], and different proof search strategies can be used to justify the design of new and different logic programming languages [54].

While the distinction between these two aspects is somewhat informal, it helps to identify and classify different concerns that arise in computational semantics. For instance, confluence and termination of reductions are crucial considerations for normalization, while unification and strategies are important for search. A key challenge of computational logic is to find means of uniting or reorganizing these apparently disjoint concerns.

An important organizational principle is structural proof theory, that is, the study of proofs as syntactic, algebraic and combinatorial objects. Formal proofs often have equivalences in their syntactic representations, leading to an important research question about *canonicity* in proofs – when are two proofs "essentially the same?" The syntactic equivalences can be used to derive normal forms for proofs that illuminate not only the proofs of a given formula, but also its entire proof search space. The celebrated *focusing* theorem of Andreoli [34] identifies one such normal form for derivations in the sequent calculus that has many important consequences both for search and for computation. The combinatorial structure of proofs can be further explored with the use of *deep inference*; in particular, deep inference allows access to simple and manifestly correct cut-elimination procedures with precise complexity bounds.

Type theory is another important organizational principle, but most popular type systems are generally designed for either search or for normalization. To give some examples, the Coq system [60] that implements the Calculus of Inductive Constructions (CIC) is designed to facilitate the expression of computational features of proofs directly as executable functional programs, but general proof search techniques for Coq are rather primitive. In contrast, the Twelf system [57] that is based on the LF type theory (a subsystem of the CIC), is based on relational specifications in canonical form (*i.e.*, without redexes) for which there are sophisticated automated reasoning systems such as meta-theoretic analysis tools, logic programming engines, and inductive theorem provers. In recent years, there has been a push towards combining search and normalization in the same type-theoretic framework. The Beluga system [58], for example, is an extension of the LF type theory with a purely computational meta-framework where operations on inductively defined LF objects can be expressed as functional programs.

The Parsifal team investigates both the search and the normalization aspects of computational specifications using the concepts, results, and insights from proof theory and type theory.

3.2. Inductive and co-inductive reasoning

The team has spent a number of years in designing a strong new logic that can be used to reason (inductively and co-inductively) on syntactic expressions containing bindings. This work is based on earlier work by McDowell, Miller, and Tiu [53] [52] [56] [61], and on more recent work by Gacek, Miller, and Nadathur [41] [40]. The Parsifal team, along with our colleagues in Minneapolis, Canberra, Singapore, and Cachan, have been building two tools that exploit the novel features of this logic. These two systems are the following.

- Abella, which is an interactive theorem prover for the full logic.
- Bedwyr, which is a model checker for the "finite" part of the logic.

We have used these systems to provide formalize reasoning of a number of complex formal systems, ranging from programming languages to the λ-calculus and π-calculus.

Since 2014, the Abella system has been extended with a number of new features. A number of new significant examples have been implemented in Abella and an extensive tutorial for it has been written [1].

3.3. Developing a foundational approach to defining proof evidence

The team is developing a framework for defining the semantics of proof evidence. With this framework, implementers of theorem provers can output proof evidence in a format of their choice: they will only need to be able to formally define that evidence's semantics. With such semantics provided, proof checkers can then check alleged proofs for correctness. Thus, anyone who needs to trust proofs from various provers can put their energies into designing trustworthy checkers that can execute the semantic specification.

In order to provide our framework with the flexibility that this ambitious plan requires, we have based our design on the most recent advances within the theory of proofs. For a number of years, various team members have been contributing to the design and theory of *focused proof systems* [35] [37] [38] [39] [43] [49] [50] and we have adopted such proof systems as the corner stone for our framework.

We have also been working for a number of years on the implementation of computational logic systems, involving, for example, both unification and backtracking search. As a result, we are also building an early and reference implementation of our semantic definitions.

3.4. Deep inference

Deep inference [44], [46] is a novel methodology for presenting deductive systems. Unlike traditional formalisms like the sequent calculus, it allows rewriting of formulas deep inside arbitrary contexts. The new freedom for designing inference rules creates a richer proof theory. For example, for systems using deep inference, we have a greater variety of normal forms for proofs than in sequent calculus or natural deduction systems. Another advantage of deep inference systems is the close relationship to category-theoretic proof theory. Due to the deep inference design one can directly read off the morphism from the derivations. There is no need for a counter-intuitive translation.

The following research problems are investigated by members of the Parsifal team:

- Find deep inference system for richer logics. This is necessary for making the proof theoretic results of deep inference accessible to applications as they are described in the previous sections of this report.
- Investigate the possibility of focusing proofs in deep inference. As described before, focusing is a way to reduce the non-determinism in proof search. However, it is well investigated only for the sequent calculus. In order to apply deep inference in proof search, we need to develop a theory of focusing for deep inference.

3.5. Proof nets, atomic flows, and combinatorial proofs

Proof nets graph-like presentations of sequent calculus proofs such that all "trivial rule permutations" are quotiented away. Ideally the notion of proof net should be independent from any syntactic formalism, but most notions of proof nets proposed in the past were formulated in terms of their relation to the sequent calculus. Consequently we could observe features like “boxes” and explicit “contraction links”. The latter appeared not only in Girard’s proof nets [42] for linear logic but also in Robinson’s proof nets [59] for classical logic. In this kind of proof nets every link in the net corresponds to a rule application in the sequent calculus.

Only recently, due to the rise of deep inference, new kinds of proof nets have been introduced that take the formula trees of the conclusions and add additional “flow-graph” information (see e.g., [48][2] leading to the notion of *atomic flow* and [45]. On one side, this gives new insights in the essence of proofs and their normalization. But on the other side, all the known correctness criteria are no longer available.

Combinatorial proofs [47] are another form syntax-independent proof presentation which separates the multiplicative from the additive behaviour of classical connectives.

The following research questions investigated by members of the Parsifal team:

- Finding (for classical and intuitionistic logic) a notion of canonical proof presentation that is deductive, i.e., can effectively be used for doing proof search.
- Studying the normalization of proofs using atomic flows and combinatorial proofs, as they simplify the normalization procedure for proofs in deep inference, and additionally allow to get new insights in the complexity of the normalization.
- Studying the size of proofs in the combinatorial proof formalism.

3.6. Cost Models and Abstract Machines for Functional Programs

In the *proof normalization* approach, computation is usually reformulated as the evaluation of functional programs, expressed as terms in a variation over the λ-calculus. Thanks to its higher-order nature, this approach provides very concise and abstract specifications. Its strength is however also its weakness: the abstraction from physical machines is pushed to a level where it is no longer clear how to measure the complexity of an algorithm.

Models like Turing machines or RAM rely on atomic computational steps and thus admit quite obvious cost models for time and space. The λ-calculus instead relies on a single non-atomic operation, β-reduction, for which costs in terms of time and space are far from evident.

Nonetheless, it turns out that the number of β-steps is a reasonable time cost model, i.e.,it is polynomially related to those of Turing machines and RAM. For the special case of *weak evaluation* (i.e., reducing only β-steps that are not under abstractions)—which is used to model functional programming languages—this is a relatively old result due to Blelloch and Greiner [36] (1995). It is only very recently (2014) that the strong case—used in the implementation models of proof assistants—has been solved by Accattoli and Dal Lago [33].

With the recent recruitment of Accattoli, the team’s research has expanded in this direction. The topics under investigations are:

1. *Complexity of Abstract Machines*. Bounding and comparing the overhead of different abstract machines for different evaluation schemas (weak/strong call-by-name/value/need λ-calculi) with respect to the cost model. The aim is the development of a complexity-aware theory of the implementation of functional programs.
2. *Reasonable Space Cost Models*. Essentially nothing is known about reasonable space cost models. It is known, however, that environment-based execution model—which are the mainstream technology for functional programs—do not provide an answer. We are exploring the use of the non-standard implementation models provided by Girard’s Geometry of Interaction to address this question.

PESTO Project-Team

3. Research Program

3.1. Modelling

Before being able to analyse and properly design security protocols, it is essential to have a model with a precise semantics of the protocols themselves, the attacker and its capabilities, as well as the properties a protocol must ensure.

Most current languages for protocol specification are quite basic and do not provide support for global state, loops, or complex data structures such as lists, or Merkle trees. As an example we may cite Hardware Security Modules that rely on a notion of *mutable global state* which does not arise in traditional protocols, see e.g. the discussion by Herzog [53].

Similarly, the properties a protocol should satisfy are generally not precisely defined, and stating the "right" definitions is often a challenging task in itself. In the case of authentication, many protocol attacks were due to the lack of a precise meaning, cf. [52]. While the case of authentication has been widely studied, the recent digitalisation of all kinds of transactions and services, introduces a plethora of new properties, including for instance anonymity in e-voting, untraceability of RFID tokens, verifiability of computations that are out-sourced, as well as sanitisation of data in social networks. We expect that many privacy and anonymity properties may be modelled as particular observational equivalences in process calculi [48], or indistinguishability between cryptographic games [3]; sanitisation of data may also rely on information-theoretic measures.

We also need to take into account that the attacker model changes. While historically the attacker was considered to control the communication network, we may nowadays argue that even (part of) the host executing the software may be compromised through, e.g., malware. This situation motivates the use of secure elements and multi-factor authentication with out-of-band channels. A typical example occurs in e-commerce: to validate an online payment a user needs to enter an additional code sent by the bank via SMS to the user's mobile phone. Such protocols require the possession of a physical device in addition to the knowledge of a password which could have been leaked on an untrusted platform. The fact that data needs to be copied by a human requires these data to be *short*, and hence amenable to brute-force attacks by an attacker or guessing.

3.2. Analysis

3.2.1. Generic proof techniques

Most automated tools for verifying security properties rely on techniques stemming from automated deduction. Often existing techniques do however not apply directly, or do not scale up due to state explosion problems. For instance, the use of Horn clause resolution techniques requires dedicated resolution methods [41] [44]. Another example is unification modulo equational theory, which is a key technique in several tools, e.g. [51]. Security protocols however require to consider particular equational theories that are not naturally studied in classical automated reasoning. Sometimes, even new concepts have been introduced. One example is the finite variant property [46], which is used in several tools, e.g., *Akiss* [44], Maude-NPA [51] and Tamarin [54]. Another example is the notion of asymmetric unification [50] which is a variant of unification used in Maude-NPA to perform important *syntactic* pruning techniques of the search space, even when reasoning modulo an equational theory. For each of these topics we need to design efficient decision procedures for a variety of equational theories.

3.2.2. Dedicated procedures and tools

We design dedicated techniques for automated protocol verification. While existing techniques for security protocol verification are efficient and have reached maturity for verification of confidentiality and authentication properties (or more generally safety properties), our goal is to go beyond these properties and the standard attacker models, verifying the properties and attacker models identified in Section 3.1 . This includes techniques that:

- can analyse *indistinguishability* properties, including for instance anonymity and unlinkability properties, but also properties stated in simulation-based (also known as universally composable) frameworks, which express the security of a protocol as an ideal (correct by design) system;
- take into account protocols that rely on a notion of *mutable global state* which does not arise in traditional protocols, but is essential when verifying tamper-resistant hardware devices, e.g., the RSA PKCS#11 standard, IBM's CCA and the trusted platform module (TPM);
- consider attacker models for protocols relying on *weak secrets* that need to be copied or remembered by a human, such as multi-factor authentication.

These goals are beyond the scope of most current analysis tools and require both theoretical advances in the area of verification, as well as the design of new efficient verification tools.

3.3. Design

Given our experience in formal analysis of security protocols, including both protocol proofs and finding of flaws, it is tempting to use our experience to design protocols with security in mind and security proofs. This part includes both provably secure design techniques, as well as the development of new protocols.

3.3.1. General design techniques

Design techniques include *composition results* that allow one to design protocols in a modular way [47], [45]. Composition results come in many flavours: they may allow one to compose protocols with different objectives, e.g. compose a key exchange protocol with a protocol that requires a shared key or rely on a protocol for secure channel establishment, compose different protocols in parallel that may re-use some key material, or compose different sessions of the same protocol.

Another area where composition is of particular importance is Service Oriented Computing, where an "orchestrator" must combine some available component services, while guaranteeing some security properties. In this context, we work on the automated synthesis of the orchestrator or monitors for enforcing the security goals. These problems require the study of new classes of automata that communicate with structured messages.

3.3.2. New protocol design

We also design new protocols. Application areas that seem of particular importance are:

- External hardware devices such as security APIs that allow for flexible key management, including key revocation, and their integration in security protocols. The security *fiasco* of the PKCS#11 standard [43], [49] witnesses the need for new protocols in this area.
- Election systems that provide strong security guarantees. We have been working (in collaboration with the Caramba team) on a prototype implementation of an e-voting system, Belenios (http://belenios.gforge.inria.fr).
- Mechanisms for publishing personal information (e.g. on social networks) in a controlled way.

PI.R2 Project-Team

3. Research Program

3.1. Proof theory and the Curry-Howard correspondence

3.1.1. Proofs as programs

Proof theory is the branch of logic devoted to the study of the structure of proofs. An essential contributor to this field is Gentsen [77] who developed in 1935 two logical formalisms that are now central to the study of proofs. These are the so-called "natural deduction", a syntax that is particularly well-suited to simulate the intuitive notion of reasoning, and the so-called "sequent calculus", a syntax with deep geometric properties that is particularly well-suited for proof automation.

Proof theory gained a remarkable importance in computer science when it became clear, after genuine observations first by Curry in 1958 [72], then by Howard and de Bruijn at the end of the 60's [89], [109], that proofs had the very same structure as programs: for instance, natural deduction proofs can be identified as typed programs of the ideal programming language known as λ-calculus.

This proofs-as-programs correspondence has been the starting point to a large spectrum of researches and results contributing to deeply connect logic and computer science. In particular, it is from this line of work that Coquand and Huet's Calculus of Constructions [69], [70] stemmed out – a formalism that is both a logic and a programming language and that is at the source of the Coq system [107].

3.1.2. Towards the calculus of constructions

The λ-calculus, defined by Church [67], is a remarkably succinct model of computation that is defined via only three constructions (abstraction of a program with respect to one of its parameters, reference to such a parameter, application of a program to an argument) and one reduction rule (substitution of the formal parameter of a program by its effective argument). The λ-calculus, which is Turing-complete, i.e. which has the same expressiveness as a Turing machine (there is for instance an encoding of numbers as functions in λ-calculus), comes with two possible semantics referred to as call-by-name and call-by-value evaluations. Of these two semantics, the first one, which is the simplest to characterise, has been deeply studied in the last decades [60].

To explain the Curry-Howard correspondence, it is important to distinguish between intuitionistic and classical logic: following Brouwer at the beginning of the 20^{th} century, classical logic is a logic that accepts the use of reasoning by contradiction while intuitionistic logic proscribes it. Then, Howard's observation is that the proofs of the intuitionistic natural deduction formalism exactly coincide with programs in the (simply typed) λ-calculus.

A major achievement has been accomplished by Martin-Löf who designed in 1971 a formalism, referred to as modern type theory, that was both a logical system and a (typed) programming language [98].

In 1985, Coquand and Huet [69], [70] in the Formel team of Inria-Rocquencourt explored an alternative approach based on Girard-Reynolds' system F [78], [102]. This formalism, called the Calculus of Constructions, served as logical foundation of the first implementation of Coq in 1984. Coq was called CoC at this time.

3.1.3. The Calculus of Inductive Constructions

The first public release of CoC dates back to 1989. The same project-team developed the programming language Caml (nowadays called OCaml and coordinated by the Gallium team) that provided the expressive and powerful concept of algebraic data types (a paragon of it being the type of lists). In CoC, it was possible to simulate algebraic data types, but only through a not-so-natural not-so-convenient encoding.

In 1989, Coquand and Paulin [71] designed an extension of the Calculus of Constructions with a generalisation of algebraic types called inductive types, leading to the Calculus of Inductive Constructions (CIC) that started to serve as a new foundation for the Coq system. This new system, which got its current definitive name Coq, was released in 1991.

In practice, the Calculus of Inductive Constructions derives its strength from being both a logic powerful enough to formalise all common mathematics (as set theory is) and an expressive richly-typed functional programming language (like ML but with a richer type system, no effects and no non-terminating functions).

3.2. The development of Coq

During 1984-2012 period, about 40 persons have contributed to the development of Coq, out of which 7 persons have contributed to bring the system to the place it was six years ago. First Thierry Coquand through his foundational theoretical ideas, then Gérard Huet who developed the first prototypes with Thierry Coquand and who headed the Coq group until 1998, then Christine Paulin who was the main actor of the system based on the CIC and who headed the development group from 1998 to 2006. On the programming side, important steps were made by Chet Murthy who raised Coq from the prototypical state to a reasonably scalable system, Jean-Christophe Filliâtre who turned to concrete the concept of a small trustful certification kernel on which an arbitrary large system can be set up, Bruno Barras and Hugo Herbelin who, among other extensions, reorganised Coq on a new smoother and more uniform basis able to support a new round of extensions for the next decade.

The development started from the Formel team at Rocquencourt but, after Christine Paulin got a position in Lyon, it spread to École Normale Supérieure de Lyon. Then, the task force there globally moved to the University of Orsay when Christine Paulin got a new position there. On the Rocquencourt side, the part of Formel involved in ML moved to the Cristal team (now Gallium) and Formel got renamed into Coq. Gérard Huet left the team and Christine Paulin started to head a Coq team bilocalised at Rocquencourt and Orsay. Gilles Dowek became the head of the team which was renamed into LogiCal. Following Gilles Dowek who got a position at École Polytechnique, LogiCal moved to the new Inria Saclay research center. It then split again, giving birth to ProVal. At the same time, the Marelle team (formerly Lemme, formerly Croap) which has been a long partner of the Formel team, invested more and more energy in the formalisation of mathematics in Coq, while contributing importantly to the development of Coq, in particular for what regards user interfaces.

After various other spreadings resulting from where the wind pushed former PhD students, the development of Coq got multi-site with the development now realised mainly by employees of Inria, the CNAM, and Paris Diderot.

In the last seven years, Hugo Herbelin and Matthieu Sozeau coordinated the development of the system, the official coordinator hat passed from Hugo to Matthieu in August 2016. The ecosystem and development model changed greatly during this period, with a move towards an entirely distributed development model, integrating contributions from all over the world. While the system had always been open-source, its development team was relatively small, well-knit and gathered regularly at Coq working groups, and many developments on Coq were still discussed only by the few interested experts.

The last years saw a big increase in opening the development to external scrutiny and contributions. This was supported by the “core” team which started moving development to the open GitHub platform (including since 2017 its bug-tracker [43] and wiki), made its development process public, starting to use public pull requests to track the work of developers, organising yearly hackatons/coding-sprints for the dissemination of expertise and developers & users meetings like the Coq Workshop and CoqPL, and, perhaps more anecdotally, retransmitting Coq working groups on a public YouTube channel.

This move was also supported by the hiring of Maxime Dénès in 2016 as an Inria research engineer (in Sophia-Antipolis), and the work of Matej Košík (2-year research engineer). Their work involved making the development process more predictable and streamlined and to provide a higher level of quality to the whole system. In se 2018, a second engineer, Vincent Laporte, was hired. Yves Bertot, Maxime Dénès and

Vincent Laporte are developing the Coq consortium, which aims to become the incarnation of the global Coq community and to offer support for our users.

Today, the development of Coq involves participants from the Inria project-teams pi.r2 (Paris), Marelle (Sophia-Antipolis), Toccata (Saclay), Gallinette (Nantes), Gallium (Paris), and Camus (Strasboug), the LIX at École Polytechnique and the CRI Mines-ParisTech. Apart from those, active collaborators include members from MPI-Saarbrucken (D. Dreyer's group), KU Leuven (B. Jacobs group), MIT CSAIL (A. Chlipala's group, which hosted an Inria/MIT engineer, and N. Zeldovich's group), the Institute for Advanced Study in Princeton (from S. Awodey, T. Coquand and V. Voevodsky's Univalent Foundations program) and Intel (M. Soegtrop). The latest released version Coq 8.8.0 had 40 contributors (counted from the start of 8.8 development) and the upcoming Coq 8.9 has 54.

On top of the developer community, there is a much wider user community, as Coq is being used in many different fields. The Software Foundations series, authored by academics from the USA, along with the reference Coq'Art book by Bertot and Castéran [61], the more advanced Certified Programming with Dependent Types book by Chlipala [66] and the recent book on the Mathematical Components library by Mahboubi, Tassi et al. provide resources for gradually learning the tool.

In the programming languages community, Coq is being taught in two summer schools, OPLSS and the DeepSpec summer school. For more mathematically inclined users, there are regular Winter Schools in Nice and in 2017 there was a school on the use of the Univalent Foundations library in Birmingham.

Since 2016, Coq also provides a central repository for Coq packages, the Coq opam archive, relying on the OCaml opam package manager and including around 250 packages contributed by users. It would be too long to make a detailed list of the uses of Coq in the wild. We only highlight four research projects relying heavily on Coq. The Mathematical Components library has its origins in the formal proof of the Four Colour Theorem and has grown to cover many areas of mathematics in Coq using the now integrated (since Coq 8.7) SSReflect proof language. The DeepSpec project is an NSF Expedition project led by A. Appel whose aim is full-stack verification of a software system, from machine-checked proofs of circuits to an operating system to a web-browser, entirely written in Coq and integrating many large projects into one. The ERC CoqHoTT project led by N. Tabareau aims to use logical tools to extend the expressive power of Coq, dealing with the univalence axiom and effects. The ERC RustBelt project led by D. Dreyer concerns the development of rigorous formal foundations for the Rust programming language, using the Iris Higher-Order Concurrent Separation Logic Framework in Coq.

We next briefly describe the main components of Coq.

3.2.1. The underlying logic and the verification kernel

The architecture adopts the so-called de Bruijn principle: the well-delimited *kernel* of Coq ensures the correctness of the proofs validated by the system. The kernel is rather stable with modifications tied to the evolution of the underlying Calculus of Inductive Constructions formalism. The kernel includes an interpreter of the programs expressible in the CIC and this interpreter exists in two flavours: a customisable lazy evaluation machine written in OCaml and a call-by-value bytecode interpreter written in C dedicated to efficient computations. The kernel also provides a module system.

3.2.2. Programming and specification languages

The concrete user language of Coq, called *Gallina*, is a high-level language built on top of the CIC. It includes a type inference algorithm, definitions by complex pattern-matching, implicit arguments, mathematical notations and various other high-level language features. This high-level language serves both for the development of programs and for the formalisation of mathematical theories. Coq also provides a large set of commands. Gallina and the commands together forms the *Vernacular* language of Coq.

3.2.3. Standard library

The standard library is written in the vernacular language of Coq. There are libraries for various arithmetical structures and various implementations of numbers (Peano numbers, implementation of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ with binary

digits, implementation of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ using machine words, axiomatisation of $\mathbb{R}$). There are libraries for lists, list of a specified length, sorts, and for various implementations of finite maps and finite sets. There are libraries on relations, sets, orders.

3.2.4. Tactics

The tactics are the methods available to conduct proofs. This includes the basic inference rules of the CIC, various advanced higher level inference rules and all the automation tactics. Regarding automation, there are tactics for solving systems of equations, for simplifying ring or field expressions, for arbitrary proof search, for semi-decidability of first-order logic and so on. There is also a powerful and popular untyped scripting language for combining tactics into more complex tactics.

Note that all tactics of Coq produce proof certificates that are checked by the kernel of Coq. As a consequence, possible bugs in proof methods do not hinder the confidence in the correctness of the Coq checker. Note also that the CIC being a programming language, tactics can have their core written (and certified) in the own language of Coq if needed.

3.2.5. Extraction

Extraction is a component of Coq that maps programs (or even computational proofs) of the CIC to functional programs (in OCaml, Scheme or Haskell). Especially, a program certified by Coq can further be extracted to a program of a full-fledged programming language then benefiting of the efficient compilation, linking tools, profiling tools, ... of the target language.

3.2.6. Documentation

Coq is a feature-rich system and requires extensive training in order to be used proficiently; current documentation includes the reference manual, the reference for the standard library, as well as tutorials, and related tooling [sphinx plugins, coqdoc]. The jsCoq tool allows writing interactive web pages were Coq programs can be embedded and executed.

3.2.7. Proof development infrastructure

Coq is used in large-scale proof developments, and provides users miscellaneous tooling to help with them: the coq_makefile and Dune build systems help with incremental proof-checking; the Coq OPAM repository contains a package index for most Coq developments; the CoqIDE, ProofGeneral, and VSCoq user interfaces are environments for proof writing; and the Coq's API does allow users to extend the system in many important ways. Among the current extensions we have QuickChik, a tool for property-based testing; STMCoq and CoqHammer integrating Coq with automated solvers; ParamCoq, providing automatic derivation of parametricity principles; MetaCoq for metaprogramming; Equations for dependently-typed programming; SerAPI, for data-centric applications; etc... This also includes the main open Coq repository living at Github.

3.3. Dependently typed programming languages

Dependently typed programming (shortly DTP) is an emerging concept referring to the diffuse and broadening tendency to develop programming languages with type systems able to express program properties finer than the usual information of simply belonging to specific data-types. The type systems of dependently-typed programming languages allow to express properties *dependent* of the input and the output of the program (for instance that a sorting program returns a list of same size as its argument). Typical examples of such languages were the Cayenne language, developed in the late 90's at Chalmers University in Sweden and the DML language developed at Boston. Since then, various new tools have been proposed, either as typed programming languages whose types embed equalities (Ωmega at Portland, ATS at Boston, ...) or as hybrid logic/programming frameworks (Agda at Chalmers University, Twelf at Carnegie, Delphin at Yale, OpTT at U. Iowa, Epigram at Nottingham, ...).

DTP contributes to a general movement leading to the fusion between logic and programming. Coq, whose language is both a logic and a programming language which moreover can be extracted to pure ML code plays a role in this movement and some frameworks combining logic and programming have been proposed on top of Coq (Concoqtion at Rice and Colorado, Ynot at Harvard, Why in the ProVal team at Inria, Iris at MPI-Saarbrucken). It also connects to Hoare logic, providing frameworks where pre- and post-conditions of programs are tied with the programs.

DTP approached from the programming language side generally benefits of a full-fledged language (e.g. supporting effects) with efficient compilation. DTP approached from the logic side generally benefits of an expressive specification logic and of proof methods so as to certify the specifications. The weakness of the approach from logic however is generally the weak support for effects or partial functions.

3.3.1. Type-checking and proof automation

In between the decidable type systems of conventional data-types based programming languages and the full expressiveness of logically undecidable formulae, an active field of research explores a spectrum of decidable or semi-decidable type systems for possible use in dependently typed programming languages. At the beginning of the spectrum, this includes, for instance, the system F's extension ML_F of the ML type system or the generalisation of abstract data types with type constraints (G.A.D.T.) such as found in the Haskell programming language. At the other side of the spectrum, one finds arbitrary complex type specification languages (e.g. that a sorting function returns a list of type "sorted list") for which more or less powerful proof automation tools exist – generally first-order ones.

3.4. Around and beyond the Curry-Howard correspondence

For two decades, the Curry-Howard correspondence has been limited to the intuitionistic case but since 1990, an important stimulus spurred on the community following Griffin's discovery that this correspondence was extensible to classical logic. The community then started to investigate unexplored potential connections between computer science and logic. One of these fields is the computational understanding of Gentzen's sequent calculus while another one is the computational content of the axiom of choice.

3.4.1. Control operators and classical logic

Indeed, a significant extension of the Curry-Howard correspondence has been obtained at the beginning of the 90's thanks to the seminal observation by Griffin [79] that some operators known as control operators were typable by the principle of double negation elimination ($\neg\neg A \Rightarrow A$), a principle that enables classical reasoning.

Control operators are used to jump from one location of a program to another. They were first considered in the 60's by Landin [96] and Reynolds [101] and started to be studied in an abstract way in the 80's by Felleisen *et al* [75], leading to Parigot's $\lambda\mu$-calculus [99], a reference calculus that is in close Curry-Howard correspondence with classical natural deduction. In this respect, control operators are fundamental pieces to establish a full connection between proofs and programs.

3.4.2. Sequent calculus

The Curry-Howard interpretation of sequent calculus started to be investigated at the beginning of the 90's. The main technicality of sequent calculus is the presence of *left introduction* inference rules, for which two kinds of interpretations are applicable. The first approach interprets left introduction rules as construction rules for a language of patterns but it does not really address the problem of the interpretation of the implication connective. The second approach, started in 1994, interprets left introduction rules as evaluation context formation rules. This line of work led in 2000 to the design by Hugo Herbelin and Pierre-Louis Curien of a symmetric calculus exhibiting deep dualities between the notion of programs and evaluation contexts and between the standard notions of call-by-name and call-by-value evaluation semantics.

3.4.3. Abstract machines

Abstract machines came as an intermediate evaluation device, between high-level programming languages and the computer microprocessor. The typical reference for call-by-value evaluation of λ-calculus is Landin's SECD machine [95] and Krivine's abstract machine for call-by-name evaluation [92], [91]. A typical abstract machine manipulates a state that consists of a program in some environment of bindings and some evaluation context traditionally encoded into a "stack".

3.4.4. Delimited control

Delimited control extends the expressiveness of control operators with effects: the fundamental result here is a completeness result by Filinski [76]: any side-effect expressible in monadic style (and this covers references, exceptions, states, dynamic bindings, ...) can be simulated in λ-calculus equipped with delimited control.

3.5. Effective higher-dimensional algebra

3.5.1. Higher-dimensional algebra

Like ordinary categories, higher-dimensional categorical structures originate in algebraic topology. Indeed, ∞-groupoids have been initially considered as a unified point of view for all the information contained in the homotopy groups of a topological space X: the *fundamental ∞-groupoid* $\Pi(X)$ of X contains the elements of X as 0-dimensional cells, continuous paths in X as 1-cells, homotopies between continuous paths as 2-cells, and so on. This point of view translates a topological problem (to determine if two given spaces X and Y are homotopically equivalent) into an algebraic problem (to determine if the fundamental groupoids $\Pi(X)$ and $\Pi(Y)$ are equivalent).

In the last decades, the importance of higher-dimensional categories has grown fast, mainly with the new trend of *categorification* that currently touches algebra and the surrounding fields of mathematics. Categorification is an informal process that consists in the study of higher-dimensional versions of known algebraic objects (such as higher Lie algebras in mathematical physics [59]) and/or of "weakened" versions of those objects, where equations hold only up to suitable equivalences (such as weak actions of monoids and groups in representation theory [74]).

The categorification process has also reached logic, with the introduction of homotopy type theory. After a preliminary result that had identified categorical structures in type theory [88], it has been observed recently that the so-called "identity types" are naturally equiped with a structure of ∞-groupoid: the 1-cells are the proofs of equality, the 2-cells are the proofs of equality between proofs of equality, and so on. The striking resemblance with the fundamental ∞-groupoid of a topological space led to the conjecture that homotopy type theory could serve as a replacement of set theory as a foundational language for different fields of mathematics, and homotopical algebra in particular.

3.5.2. Higher-dimensional rewriting

Higher-dimensional categories are algebraic structures that contain, in essence, computational aspects. This has been recognised by Street [106], and independently by Burroni [64], when they have introduced the concept of *computad* or *polygraph* as combinatorial descriptions of higher categories. Those are directed presentations of higher-dimensional categories, generalising word and term rewriting systems.

In the recent years, the algebraic structure of polygraph has led to a new theory of rewriting, called *higher-dimensional rewriting*, as a unifying point of view for usual rewriting paradigms, namely abstract, word and term rewriting [93], [97], [80], [81], and beyond: Petri nets [83] and formal proofs of classical and linear logic have been expressed in this framework [82]. Higher-dimensional rewriting has developed its own methods to analyse computational properties of polygraphs, using in particular algebraic tools such as derivations to prove termination, which in turn led to new tools for complexity analysis [62].

3.5.3. Squier theory

The homotopical properties of higher categories, as studied in mathematics, are in fact deeply related to the computational properties of their polygraphic presentations. This connection has its roots in a tradition of using rewriting-like methods in algebra, and more specifically in the works of Anick [57] and Squier [105], [104]: Squier has proved that, if a monoid M can be presented by a *finite*, *terminating* and *confluent* rewriting system, then its third integral homology group $H_3(M, \mathbb{Z})$ is finitely generated and the monoid M has *finite derivation type* (a property of homotopical nature). This allowed him to conclude that finite convergent rewriting systems were not a universal solution to decide the word problem of finitely generated monoids. Since then, Yves Guiraud and Philippe Malbos have shown that this connection was part of a deeper unified theory when formulated in the higher-dimensional setting [14], [15], [85], [86], [87].

In particular, the computational content of Squier's proof has led to a constructive methodology to produce, from a convergent presentation, *coherent presentations* and *polygraphic resolutions* of algebraic structures, such as monoids [14] and algebras [32]. A coherent presentation of a monoid M is a 3-dimensional combinatorial object that contains not only a presentation of M (generators and relations), but also higher-dimensional cells, corresponding each to two fundamentally different proofs of the same equality: this is, in essence, the same as the proofs of equality of proofs of equality in homotopy type theory. When this process of "unfolding" proofs of equalities is pursued in every dimension, one gets a polygraphic resolution of the starting monoid M. This object has the following desirable qualities: it is free and homotopically equivalent to M (in the canonical model structure of higher categories [94], [58]). A polygraphic resolution of an algebraic object X is a faithful formalisation of X on which one can perform computations, such as homotopical or homological invariants of X. In particular, this has led to new algorithms and proofs in representation theory [11], and in homological algebra [84][32].

POLSYS Project-Team

3. Research Program

3.1. Introduction

Polynomial system solving is a fundamental problem in Computer Algebra with many applications in cryptography, robotics, biology, error correcting codes, signal theory, ... Among all available methods for solving polynomial systems, computation of Gröbner bases remains one of the most powerful and versatile method since it can be applied in the continuous case (rational coefficients) as well as in the discrete case (finite fields). Gröbner bases are also building blocks for higher level algorithms that compute real sample points in the solution set of polynomial systems, decide connectivity queries and quantifier elimination over the reals. The major challenge facing the designer or the user of such algorithms is the intrinsic exponential behaviour of the complexity for computing Gröbner bases. The current proposal is an attempt to tackle these issues in a number of different ways: improve the efficiency of the fundamental algorithms (even when the complexity is exponential), develop high performance implementation exploiting parallel computers, and investigate new classes of structured algebraic problems where the complexity drops to polynomial time.

3.2. Fundamental Algorithms and Structured Systems

Participants: Jérémy Berthomieu, Jean-Charles Faugère, Mohab Safey El Din, Elias Tsigaridas, Dongming Wang, Matías Bender, Thi Xuan Vu.

Efficient algorithms F_4/F_5[0] for computing the Gröbner basis of a polynomial system rely heavily on a connection with linear algebra. Indeed, these algorithms reduce the Gröbner basis computation to a sequence of Gaussian eliminations on several submatrices of the so-called Macaulay matrix in some degree. Thus, we expect to improve the existing algorithms by
(i) developing dedicated linear algebra routines performing the Gaussian elimination steps: this is precisely the objective 2 described below;
(ii) generating smaller or simpler matrices to which we will apply Gaussian elimination.
We describe here our goals for the latter problem. First, we focus on algorithms for computing a Gröbner basis of *general polynomial systems*. Next, we present our goals on the development of dedicated algorithms for computing Gröbner bases of *structured polynomial systems* which arise in various applications.

Algorithms for general systems. Several degrees of freedom are available to the designer of a Gröbner basis algorithm to generate the matrices occurring during the computation. For instance, it would be desirable to obtain matrices which would be almost triangular or very sparse. Such a goal can be achieved by considering various interpretations of the F_5 algorithm with respect to different monomial orderings. To address this problem, the tight complexity results obtained for F_5 will be used to help in the design of such a general algorithm. To illustrate this point, consider the important problem of solving boolean polynomial systems; it might be interesting to preserve the sparsity of the original equations and, at the same time, using the fact that overdetermined systems are much easier to solve.

Algorithms dedicated to *structured* polynomial systems. A complementary approach is to exploit the structure of the input polynomials to design specific algorithms. Very often, problems coming from applications are not random but are highly structured. The specific nature of these systems may vary a lot: some polynomial systems can be sparse (when the number of terms in each equation is low), overdetermined (the number of the equations is larger than the number of variables), invariants by the action of some finite groups, multi-linear (each equation is linear w.r.t. to one block of variables) or more generally multihomogeneous. In each case, the ultimate goal is to identify large classes of problems whose theoretical/practical complexity drops and to propose in each case dedicated algorithms.

[0] J.-C. Faugère. *A new efficient algorithm for computing Gröbner bases without reduction to zero (F5).* In Proceedings of ISSAC '02, pages 75-83, New York, NY, USA, 2002. ACM.

3.3. Solving Systems over the Reals and Applications.

Participants: Mohab Safey El Din, Elias Tsigaridas, Daniel Lazard, Thi Xuan Vu.

We shall develop algorithms for solving polynomial systems over complex/real numbers. Again, the goal is to extend significantly the range of reachable applications using algebraic techniques based on Gröbner bases and dedicated linear algebra routines. Targeted application domains are global optimization problems, stability of dynamical systems (e.g. arising in biology or in control theory) and theorem proving in computational geometry.

The following functionalities shall be requested by the end-users:
(i) deciding the emptiness of the real solution set of systems of polynomial equations and inequalities,
(ii) quantifier elimination over the reals or complex numbers,
(iii) answering connectivity queries for such real solution sets.
We will focus on these functionalities.

We will develop algorithms based on the so-called critical point method to tackle systems of equations and inequalities (problem *(i)*) . These techniques are based on solving 0-dimensional polynomial systems encoding "critical points" which are defined by the vanishing of minors of Jacobian matrices (with polynomial entries). Since these systems are highly structured, the expected results of Objective 1 and 2 may allow us to obtain dramatic improvements in the computation of Gröbner bases of such polynomial systems. This will be the foundation of practically fast implementations (based on singly exponential algorithms) outperforming the current ones based on the historical Cylindrical Algebraic Decomposition (CAD) algorithm (whose complexity is doubly exponential in the number of variables). We will also develop algorithms and implementations that allow us to analyze, at least locally, the topology of solution sets in some specific situations. A long-term goal is obviously to obtain an analysis of the global topology.

3.4. Low level implementation and Dedicated Algebraic Computation and Linear Algebra.

Participants: Jean-Charles Faugère, Mohab Safey El Din, Elias Tsigaridas, Olive Chakraborty, Jocelyn Ryckeghem.

Here, the primary objective is to focus on *dedicated* algorithms and software for the linear algebra steps in Gröbner bases computations and for problems arising in Number Theory. As explained above, linear algebra is a key step in the process of computing efficiently Gröbner bases. It is then natural to develop specific linear algebra algorithms and implementations to further strengthen the existing software. Conversely, Gröbner bases computation is often a key ingredient in higher level algorithms from Algebraic Number Theory. In these cases, the algebraic problems are very particular and specific. Hence dedicated Gröbner bases algorithms and implementations would provide a better efficiency.

Dedicated linear algebra tools. The FGB library is an efficient one for Gröbner bases computations which can be used, for instance, via MAPLE. However, the library is sequential. A goal of the project is to extend its efficiency to new trend parallel architectures such as clusters of multi-processor systems in order to tackle a broader class of problems for several applications. Consequently, our first aim is to provide a durable, long term software solution, which will be the successor of the existing FGB library. To achieve this goal, we will first develop a high performance linear algebra package (under the LGPL license). This could be organized in the form of a collaborative project between the members of the team. The objective is not to develop a general library similar to the LINBOX[0] project but to propose a dedicated linear algebra package taking into account the specific properties of the matrices generated by the Gröbner bases algorithms. Indeed these matrices are sparse (the actual sparsity depends strongly on the application), almost block triangular and not necessarily of full rank. Moreover, most of the pivots are known at the beginning of the computation. In practice, such matrices are huge (more than 10^6 columns) but taking into account their shape may allow us to speed up the computations by one or several orders of magnitude. A variant of a Gaussian elimination algorithm together

[0] http://www.linalg.org/

with a corresponding C implementation has been presented. The main peculiarity is the order in which the operations are performed. This will be the kernel of the new linear algebra library that will be developed.

Fast linear algebra packages would also benefit to the transformation of a Gröbner basis of a zero–dimensional ideal with respect to a given monomial ordering into a Gröbner basis with respect to another ordering. In the generic case at least, the change of ordering is equivalent to the computation of the minimal polynomial of a so-called multiplication matrix. By taking into account the sparsity of this matrix, the computation of the Gröbner basis can be done more efficiently using a variant of the Wiedemann algorithm. Hence, our goal is also to obtain a dedicated high performance library for transforming (i.e. change ordering) Gröbner bases.

Dedicated algebraic tools for Algebraic Number Theory. Recent results in Algebraic Number Theory tend to show that the computation of Gröbner basis is a key step toward the resolution of difficult problems in this domain [0]. Using existing resolution methods is simply not enough to solve relevant problems. The main algorithmic bottleneck to overcome is to adapt the Gröbner basis computation step to the specific problems. Typically, problems coming from Algebraic Number Theory usually have a lot of symmetries or the input systems are very structured. This is the case, in particular, for problems coming from the algorithmic theory of Abelian varieties over finite fields [0] where the objects are represented by polynomial system and are endowed with intrinsic group actions. The main goal here is to provide dedicated algebraic resolution algorithms and implementations for solving such problems. We do not restrict our focus on problems in positive characteristic. For instance, tower of algebraic fields can be viewed as triangular sets; more generally, related problems (e.g. effective Galois theory) which can be represented by polynomial systems will receive our attention. This is motivated by the fact that, for example, computing small integer solutions of Diophantine polynomial systems in connection with Coppersmith's method would also gain in efficiency by using a dedicated Gröbner bases computations step.

3.5. Solving Systems in Finite Fields, Applications in Cryptology and Algebraic Number Theory.

Participants: Jérémy Berthomieu, Jean-Charles Faugère, Ludovic Perret, Olive Chakraborty, Nagardjun Chinthamani, Solane El Hirch, Jocelyn Ryckeghem.

Here, we focus on solving polynomial systems over finite fields (i.e. the discrete case) and the corresponding applications (Cryptology, Error Correcting Codes, ...). Obviously this objective can be seen as an application of the results of the two previous objectives. However, we would like to emphasize that it is also the source of new theoretical problems and practical challenges. We propose to develop a systematic use of *structured systems* in *algebraic cryptanalysis*.
(i) So far, breaking a cryptosystem using algebraic techniques could be summarized as modeling the problem by algebraic equations and then computing a, usually, time consuming Gröbner basis. A new trend in this field is to require a theoretical complexity analysis. This is needed to explain the behavior of the attack but also to help the designers of new cryptosystems to propose actual secure parameters.
(ii) To assess the security of several cryptosystems in symmetric cryptography (block ciphers, hash functions, ...), a major difficulty is the size of the systems involved for this type of attack. More specifically, the bottleneck is the size of the linear algebra problems generated during a Gröbner basis computation.

We propose to develop a systematic use of *structured systems* in *algebraic cryptanalysis*.

[0] P. Gaudry, *Index calculus for abelian varieties of small dimension and the elliptic curve discrete logarithm problem*, Journal of Symbolic Computation 44,12 (2009) pp. 1690-1702
[0] e.g. point counting, discrete logarithm, isogeny.

The first objective is to build on the recent breakthrough in attacking McEliece's cryptosystem: it is the first structural weakness observed on one of the oldest public key cryptosystem. We plan to develop a well founded framework for assessing the security of public key cryptosystems based on coding theory from the algebraic cryptanalysis point of view. The answer to this issue is strongly related to the complexity of solving bihomogeneous systems (of bidegree $(1, d)$). We also plan to use the recently gained understanding on the complexity of structured systems in other areas of cryptography. For instance, the MinRank problem – which can be modeled as an overdetermined system of bilinear equations – is at the heart of the structural attack proposed by Kipnis and Shamir against HFE (one of the most well known multivariate public cryptosystem). The same family of structured systems arises in the algebraic cryptanalysis of the Discrete Logarithmic Problem (DLP) over curves (defined over some finite fields). More precisely, some bilinear systems appear in the polynomial modeling the points decomposition problem. Moreover, in this context, a natural group action can also be used during the resolution of the considered polynomial system.

Dedicated tools for linear algebra problems generated during the Gröbner basis computation will be used in algebraic cryptanalysis. The promise of considerable algebraic computing power beyond the capability of any standard computer algebra system will enable us to attack various cryptosystems or at least to propose accurate secure parameters for several important cryptosystems. Dedicated linear tools are thus needed to tackle these problems. From a theoretical perspective, we plan to further improve the theoretical complexity of the hybrid method and to investigate the problem of solving polynomial systems with noise, i.e. some equations of the system are incorrect. The hybrid method is a specific method for solving polynomial systems over finite fields. The idea is to mix exhaustive search and Gröbner basis computation to take advantage of the over-determinacy of the resulting systems.

Polynomial system with noise is currently emerging as a problem of major interest in cryptography. This problem is a key to further develop new applications of algebraic techniques; typically in side-channel and statistical attacks. We also emphasize that recently a connection has been established between several classical lattice problems (such as the Shortest Vector Problem), polynomial system solving and polynomial systems with noise. The main issue is that there is no sound algorithmic and theoretical framework for solving polynomial systems with noise. The development of such framework is a long-term objective.

PRIVATICS Project-Team (section vide)

PROSECCO Project-Team

3. Research Program

3.1. Symbolic verification of cryptographic applications

Despite decades of experience, designing and implementing cryptographic applications remains dangerously error-prone, even for experts. This is partly because cryptographic security is an inherently hard problem, and partly because automated verification tools require carefully-crafted inputs and are not widely applicable. To take just the example of TLS, a widely-deployed and well-studied cryptographic protocol designed, implemented, and verified by security experts, the lack of a formal proof about all its details has regularly led to the discovery of major attacks (including several in PROSECCO) on both the protocol and its implementations, after many years of unsuspecting use.

As a result, the automated verification for cryptographic applications is an active area of research, with a wide variety of tools being employed for verifying different kinds of applications.

In previous work, we have developed the following three approaches:

- ProVerif: a symbolic prover for cryptographic protocol models
- Tookan: an attack-finder for PKCS#11 hardware security devices
- F*: a new language that enables the verification of cryptographic applications

3.1.1. Verifying cryptographic protocols with ProVerif

Given a model of a cryptographic protocol, the problem is to verify that an active attacker, possibly with access to some cryptographic keys but unable to guess other secrets, cannot thwart security goals such as authentication and secrecy [53]; it has motivated a serious research effort on the formal analysis of cryptographic protocols, starting with [49] and eventually leading to effective verification tools, such as our tool ProVerif.

To use ProVerif, one encodes a protocol model in a formal language, called the applied pi-calculus, and ProVerif abstracts it to a set of generalized Horn clauses. This abstraction is a small approximation: it just ignores the number of repetitions of each action, so ProVerif is still very precise, more precise than, say, tree automata-based techniques. The price to pay for this precision is that ProVerif does not always terminate; however, it terminates in most cases in practice, and it always terminates on the interesting class of *tagged protocols* [44]. ProVerif can handle a wide variety of cryptographic primitives, defined by rewrite rules or by some equations, and prove a wide variety of security properties: secrecy [42], [30], correspondences (including authentication) [43], and observational equivalences [41]. Observational equivalence means that an adversary cannot distinguish two processes (protocols); equivalences can be used to formalize a wide range of properties, but they are particularly difficult to prove. Even if the class of equivalences that ProVerif can prove is limited to equivalences between processes that differ only by the terms they contain, these equivalences are useful in practice and ProVerif has long been the only tool that proves equivalences for an unbounded number of sessions. (Maude-NPA in 2014 and Tamarin in 2015 adopted ProVerif's approach to proving equivalences.)

Using ProVerif, it is now possible to verify large parts of industrial-strength protocols, such as TLS [36], Signal [51], JFK [31], and Web Services Security [40], against powerful adversaries that can run an unlimited number of protocol sessions, for strong security properties expressed as correspondence queries or equivalence assertions. ProVerif is used by many teams at the international level, and has been used in more than 120 research papers (references available at http://proverif.inria.fr/proverif-users.html).

3.1.2. Verifying security APIs using Tookan

Security application programming interfaces (APIs) are interfaces that provide access to functionality while also enforcing a security policy, so that even if a malicious program makes calls to the interface, certain security properties will continue to hold. They are used, for example, by cryptographic devices such as smartcards and Hardware Security Modules (HSMs) to manage keys and provide access to cryptographic functions whilst keeping the keys secure. Like security protocols, their design is security critical and very difficult to get right. Hence formal techniques have been adapted from security protocols to security APIs.

The most widely used standard for cryptographic APIs is RSA PKCS#11, ubiquitous in devices from smartcards to HSMs. A 2003 paper highlighted possible flaws in PKCS#11 [46], results which were extended by formal analysis work using a Dolev-Yao style model of the standard [47]. However at this point it was not clear to what extent these flaws affected real commercial devices, since the standard is underspecified and can be implemented in many different ways. The Tookan tool, developed by Steel in collaboration with Bortolozzo, Centenaro and Focardi, was designed to address this problem. Tookan can reverse engineer the particular configuration of PKCS#11 used by a device under test by sending a carefully designed series of PKCS#11 commands and observing the return codes. These codes are used to instantiate a Dolev-Yao model of the device's API. This model can then be searched using a security protocol model checking tool to find attacks. If an attack is found, Tookan converts the trace from the model checker into the sequence of PKCS#11 queries needed to make the attack and executes the commands directly on the device. Results obtained by Tookan are remarkable: of 18 commercially available PKCS#11 devices tested, 10 were found to be susceptible to at least one attack.

3.1.3. Verifying cryptographic applications using F*

Verifying the implementation of a protocol has traditionally been considered much harder than verifying its model. This is mainly because implementations have to consider real-world details of the protocol, such as message formats [55], that models typically ignore. So even a protocol has been proved secure in theory, its implementation may be buggy and insecure. However, with recent advances in both program verification and symbolic protocol verification tools, it has become possible to verify fully functional protocol implementations in the symbolic model. One approach is to extract a symbolic protocol model from an implementation and then verify the model, say, using ProVerif. This approach has been quite successful, yielding a verified implementation of TLS in F# [39]. However, the generated models are typically quite large and whole-program symbolic verification does not scale very well.

An alternate approach is to develop a verification method directly for implementation code, using well-known program verification techniques. Our current focus is on designing and implementing the programming language F* [57], [34], [52], in collaboration with Microsoft Research. F* (pronounced F star) is an ML-like functional programming language aimed at program verification. Its type system includes polymorphism, dependent types, monadic effects, refinement types, and a weakest precondition calculus. Together, these features allow expressing precise and compact specifications for programs, including functional correctness and security properties. The F* type-checker aims to prove that programs meet their specifications using a combination of SMT solving and interactive proofs[23]. Programs written in F* can be translated to efficient OCaml, F#, or C for execution [54]. The main ongoing use case of F* is building a verified, drop-in replacement for the whole HTTPS stack in Project Everest [37] (a larger collaboration with Microsoft Research). This includes a verified implementation of TLS 1.2 and 1.3 [38] and of the underlying cryptographic primitives [58].

3.2. Computational verification of cryptographic applications

Proofs done by cryptographers in the computational model are mostly manual. Our goal is to provide computer support to build or verify these proofs. In order to reach this goal, we have designed the automatic tool CryptoVerif, which generates proofs by sequences of games. We already applied it to important protocols such as TLS [36] and Signal [51] but more work is still needed in order to develop this approach, so that it is easier to apply to more protocols. We also design and implement techniques for proving implementations

of protocols secure in the computational model. In particular, CryptoVerif can generate implementations from CryptoVerif specifications that have been proved secure [45]. We plan to continue working on this approach.

A different approach is to directly verify cryptographic applications in the computational model by typing. A recent work [50] shows how to use refinement typechecking in F7 to prove computational security for protocol implementations. In this method, henceforth referred to as computational F7, typechecking is used as the main step to justify a classic game-hopping proof of computational security. The correctness of this method is based on a probabilistic semantics of F# programs and crucially relies on uses of type abstraction and parametricity to establish strong security properties, such as indistinguishability.

In principle, the two approaches, typechecking and game-based proofs, are complementary. Understanding how to combine these approaches remains an open and active topic of research.

An alternative to direct computation proofs is to identify the cryptographic assumptions under which symbolic proofs, which are typically easier to derive automatically, can be mapped to computational proofs. This line of research is sometimes called computational soundness and the extent of its applicability to real-world cryptographic protocols is an active area of investigation.

3.3. F*: A Higher-Order Effectful Language for Program Verification

F* [57], [34] is a verification system for effectful programs developed collaboratively by Inria and Microsoft Research. It puts together the automation of an SMT-backed deductive verification tool with the expressive power of a proof assistant based on dependent types. After verification, F* programs can be extracted to efficient OCaml, F#, or C code [54]. This enables verifying the functional correctness and security of realistic applications. F*'s type system includes dependent types, monadic effects, refinement types, and a weakest precondition calculus. Together, these features allow expressing precise and compact specifications for programs, including functional correctness and security properties. The F* type-checker aims to prove that programs meet their specifications using a combination of SMT solving and interactive proofs. The main ongoing use case of F* is building a verified, drop-in replacement for the whole HTTPS stack in Project Everest. This includes verified implementations of TLS 1.2 and 1.3 [38] and of the underlying cryptographic primitives [58].

3.4. Efficient Formally Secure Compilers to a Tagged Architecture

Severe low-level vulnerabilities abound in today's computer systems, allowing cyber-attackers to remotely gain full control. This happens in big part because our programming languages, compilers, and architectures were designed in an era of scarce hardware resources and too often trade off security for efficiency. The semantics of mainstream low-level languages like C is inherently insecure, and even for safer languages, establishing security with respect to a high-level semantics does not guarantee the absence of low-level attacks. Secure compilation using the coarse-grained protection mechanisms provided by mainstream hardware architectures would be too inefficient for most practical scenarios.

We aim to leverage emerging hardware capabilities for fine-grained protection to build the first, efficient secure compilation chains for realistic low-level programming languages (the C language, and Low* a safe subset of C embedded in F* for verification [54]). These compilation chains will provide a secure semantics for all programs and will ensure that high-level abstractions cannot be violated even when interacting with untrusted low-level code. To achieve this level of security without sacrificing efficiency, our secure compilation chains target a tagged architecture [35], which associates a metadata tag to each word and efficiently propagates and checks tags according to software-defined rules. We hope to experimentally evaluate and carefully optimize the efficiency of our secure compilation chains on realistic workloads and standard benchmark suites. We are also using property-based testing and formal verification to provide high confidence that our compilation chains are indeed secure. Formally, we are constructing machine-checked proofs of a new security criterion we call robustly safe compilation, which is defined as the preservation of safety properties even against an adversarial context [32], [33]. This strong criterion complements compiler correctness and ensures that no machine-code attacker can do more harm to securely compiled components than a component already could with respect to a secure source-level semantics.

3.5. Provably secure web applications

Web applications are fast becoming the dominant programming platform for new software, probably because they offer a quick and easy way for developers to deploy and sell their *app*s to a large number of customers. Third-party web-based apps for Facebook, Apple, and Google, already number in the hundreds of thousands and are likely to grow in number. Many of these applications store and manage private user data, such as health information, credit card data, and GPS locations. To protect this data, applications tend to use an ad hoc combination of cryptographic primitives and protocols. Since designing cryptographic applications is easy to get wrong even for experts, we believe this is an opportune moment to develop security libraries and verification techniques to help web application programmers.

As a typical example, consider commercial password managers, such as LastPass, RoboForm, and 1Password. They are implemented as browser-based web applications that, for a monthly fee, offer to store a user's passwords securely on the web and synchronize them across all of the user's computers and smartphones. The passwords are encrypted using a master password (known only to the user) and stored in the cloud. Hence, no-one except the user should ever be able to read her passwords. When the user visits a web page that has a login form, the password manager asks the user to decrypt her password for this website and automatically fills in the login form. Hence, the user no longer has to remember passwords (except her master password) and all her passwords are available on every computer she uses.

Password managers are available as browser extensions for mainstream browsers such as Firefox, Chrome, and Internet Explorer, and as downloadable apps for Android and Apple phones. So, seen as a distributed application, each password manager application consists of a web service (written in PHP or Java), some number of browser extensions (written in JavaScript), and some smartphone apps (written in Java or Objective C). Each of these components uses a different cryptographic library to encrypt and decrypt password data. How do we verify the correctness of all these components?

We propose three approaches. For client-side web applications and browser extensions written in JavaScript, we propose to build a static and dynamic program analysis framework to verify security invariants. To this end, we have developed two security-oriented type systems for JavaScript, Defensive JavaScript [48] and TS* [56], and used them to guarantee security properties for a number of JavaScript applications. For Android smartphone apps and web services written in Java, we propose to develop annotated JML cryptography libraries that can be used with static analysis tools like ESC/Java to verify the security of application code. For clients and web services written in F# for the .NET platform, we propose to use F* to verify their correctness. We also propose to translate verified F* web applications to JavaScript via a verified compiler that preserves the semantics of F* programs in JavaScript.

3.6. Design and Verification of next-generation protocols: identity, blockchains, and messaging

Building on our work on verifying and re-designing pre-existing protocols like TLS and Web Security in general, with the resources provided by the NEXTLEAP project, we are working on both designing and verifying new protocols in rapidly emerging areas like identity, blockchains, and secure messaging. These are all areas where existing protocols, such as the heavily used OAuth protocol, are in need of considerable re-design in order to maintain privacy and security properties. Other emerging areas, such as blockchains and secure messaging, can have modifications to existing pre-standard proposals or even a complete 'clean slate' design. As shown by Prosecco's work, newer standards, such as IETF OAuth, W3C Web Crypto, and W3C Web Authentication API, can have vulnerabilities fixed before standardization is complete and heavily deployed. We hope that the tools used by Prosecco can shape the design of new protocols even before they are shipped to standards bodies. We have seen considerable progress in identity with the UnlimitID design and with messaging via the IETF MLS effort, with new work on blockchain technology underway.

SECRET Project-Team

3. Research Program

3.1. Scientific foundations

Our approach relies on a competence whose impact is much wider than cryptology. Our tools come from information theory, discrete mathematics, probabilities, algorithmics, quantum physics... Most of our work mixes fundamental aspects (study of mathematical objects) and practical aspects (cryptanalysis, design of algorithms, implementations). Our research is mainly driven by the belief that discrete mathematics and algorithmics of finite structures form the scientific core of (algorithmic) data protection.

3.2. Symmetric cryptology

Symmetric techniques are widely used because they are the only ones that can achieve some major features such as high-speed or low-cost encryption, fast authentication, and efficient hashing. It is a very active research area which is stimulated by a pressing industrial demand. The process which has led to the new block cipher standard AES in 2001 was the outcome of a decade of research in symmetric cryptography, where new attacks have been proposed, analyzed and then thwarted by some appropriate designs. However, even if its security has not been challenged so far, it clearly appears that the AES cannot serve as a Swiss knife in all environments. In particular an important challenge raised by several new applications is the design of symmetric encryption schemes with some additional properties compared to the AES, either in terms of implementation performance (low-cost hardware implementation, low latency, resistance against side-channel attacks...) or in terms of functionalities (like authenticated encryption). The past decade has then been characterized by a multiplicity of new proposals. This proliferation of symmetric primitives has been amplified by several public competitions (eSTREAM, SHA-3, CAESAR...) which have encouraged innovative constructions and promising but unconventional designs. We are then facing up to a very new situation where implementers need to make informed choices among more than 40 lightweight block ciphers [0] or 57 new authenticated-encryption schemes [0]. Evaluating the security of all these proposals has then become a primordial task which requires the attention of the community.

In this context we believe that the cryptanalysis effort cannot scale up without an in-depth study of the involved algorithms. Indeed most attacks are described as ad-hoc techniques dedicated to a particular cipher. To determine whether they apply to some other primitives, it is then crucial to formalize them in a general setting. Our approach relies on the idea that a unified description of generic attacks (in the sense that they apply to a large class of primitives) is the only methodology for a precise evaluation of the resistance of all these new proposals, and of their security margins. In particular, such a work prevents misleading analyses based on wrong estimations of the complexity or on non-optimized algorithms. It also provides security criteria which enable designers to guarantee that their primitive resists some families of attacks. The main challenge is to provide a generic description which captures most possible optimizations of the attack.

3.3. Code-based cryptography

Public-key cryptography is one of the key tools for providing network security (SSL, e-commerce, e-banking...). The security of nearly all public-key schemes used today relies on the presumed difficulty of two problems, namely factorization of large integers or computing the discrete logarithm over various groups. The hardness of those problems was questioned in 1994 [0] when Shor showed that a quantum computer could solve them efficiently. Though large enough quantum computers that would be able to threaten the

[0] 35 are described on https://www.cryptolux.org/index.php/Lightweight_Block_Ciphers.
[0] see http://competitions.cr.yp.to/caesar-submissions.html
[0] P. Shor, *Algorithms for quantum computation: Discrete logarithms and factoring*, FOCS 1994.

existing cryptosystems do not exist yet, the cryptographic research community has to get ready and has to prepare alternatives. This line of work is usually referred to as *post-quantum cryptography*. This has become a prominent research field. Most notably, an international call for post-quantum primitives [0] has been launched by the NIST, with a submission deadline in November 2017.

The research of the project-team in this field is focused on the design and cryptanalysis of cryptosystems making use of coding theory. Code-based cryptography is one the main techniques for post-quantum cryptography (together with lattice-based, multivariate, or hash-based cryptography).

3.4. Quantum information

The field of quantum information and computation aims at exploiting the laws of quantum physics to manipulate information in radically novel ways. There are two main applications:

(i) quantum computing, that offers the promise of solving some problems that seem to be intractable for classical computers such as for instance factorization or solving the discrete logarithm problem;

(ii) quantum cryptography, which provides new ways to exchange data in a provably secure fashion. For instance it allows key distribution by using an authenticated channel and quantum communication over an unreliable channel with information-theoretic security, in the sense that its security can be proven rigorously by using only the laws of quantum physics, even with all-powerful adversaries.

Our team deals with quantum coding theoretic issues related to building a large quantum computer and with quantum cryptography. The first part builds upon our expertise in classical coding theory whereas the second axis focuses on obtaining security proofs for quantum protocols or on devising quantum cryptographic protocols (and more generally quantum protocols related to cryptography). A close relationship with partners working in the whole area of quantum information processing in the Parisian region has also been developed through our participation to the Fédération de Recherche “PCQC” (Paris Centre for Quantum Computing).

[0] http://csrc.nist.gov/groups/ST/post-quantum-crypto/

SPADES Project-Team

3. Research Program

3.1. Introduction

The SPADES research program is organized around three main themes, *Design and Programming Models*, *Certified real-time programming*, and *Fault management and causal analysis*, that seek to answer the three key questions identified in Section 2.1 . We plan to do so by developing and/or building on programming languages and techniques based on formal methods and formal semantics (hence the use of *"sound programming"* in the project-team title). In particular, we seek to support design where correctness is obtained by construction, relying on proven tools and verified constructs, with programming languages and programming abstractions designed with verification in mind.

3.2. Design and Programming Models

Work on this theme aims to develop models , languages and tools to support a "correct-by-construction" approach to the development of embedded systems.

On the programming side, we focus on the definition of domain specific programming models and languages supporting static analyses for the computation of precise resource bounds for program executions. We propose dataflow models supporting dynamicity while enjoying effective analyses. In particular, we study parametric extensions where properties such as liveness and boundedness remain statically analyzable.

On the design side, we focus on the definition of component-based models for software architectures combining distribution, dynamicity, real-time and fault-tolerant aspects. Component-based construction has long been advocated as a key approach to the "correct-by-construction" design of complex embedded systems [55]. Witness component-based toolsets such as PTOLEMY [47], BIP [38], or the modular architecture frameworks used, for instance, in the automotive industry (AUTOSAR) [30]. For building large, complex systems, a key feature of component-based construction is the ability to associate with components a set of *contracts*, which can be understood as rich behavioral types that can be composed and verified to guarantee a component assemblage will meet desired properties.

Formal models for component-based design are an active area of research. However, we are still missing a comprehensive formal model and its associated behavioral theory able to deal *at the same time* with different forms of composition, dynamic component structures, and quantitative constraints (such as timing, fault-tolerance, or energy consumption).

We plan to develop our component theory by progressing on two fronts: a semantical framework and domain-specific programming models. The work on the semantical framework should, in the longer term, provide abstract mathematical models for the more operational and linguistic analysis afforded by component calculi. Our work on component theory will find its application in the development of a COQ-based toolchain for the certified design and construction of dependable embedded systems, which constitutes our first main objective for this axis.

3.3. Certified Real-Time Programming

Programming real-time systems (*i.e.*, systems whose correct behavior depends on meeting timing constraints) requires appropriate languages (as exemplified by the family of synchronous languages [40]), but also the support of efficient scheduling policies, execution time and schedulability analyses to guarantee real-time constraints (*e.g.*, deadlines) while making the most effective use of available (processing, memory, or networking) resources. Schedulability analysis involves analyzing the worst-case behavior of real-time tasks under a given scheduling algorithm and is crucial to guarantee that time constraints are met in any possible execution of the system. Reactive programming and real-time scheduling and schedulability for multiprocessor

systems are old subjects, but they are nowhere as mature as their uniprocessor counterparts, and still feature a number of open research questions [36], [45], in particular in relation with mixed criticality systems. The main goal in this theme is to address several of these open questions.

We intend to focus on two issues: multicriteria scheduling on multiprocessors, and schedulability analysis for real-time multiprocessor systems. Beyond real-time aspects, multiprocessor environments, and multicore ones in particular, are subject to several constraints *in conjunction*, typically involving real-time, reliability and energy-efficiency constraints, making the scheduling problem more complex for both the offline and the online cases. Schedulability analysis for multiprocessor systems, in particular for systems with mixed criticality tasks, is still very much an open research area.

Distributed reactive programming is rightly singled out as a major open issue in the recent, but heavily biased (it essentially ignores recent research in synchronous and dataflow programming), survey by Bainomugisha et al. [36]. For our part, we intend to focus on devising synchronous programming languages for distributed systems and precision-timed architectures.

3.4. Fault Management and Causal Analysis

Managing faults is a clear and present necessity in networked embedded systems. At the hardware level, modern multicore architectures are manufactured using inherently unreliable technologies [41], [51]. The evolution of embedded systems towards increasingly distributed architectures highlighted in the introductory section means that dealing with partial failures, as in Web-based distributed systems, becomes an important issue.

In this axis we intend to address the question of *how to cope with faults and failures in embedded systems?*. We will tackle this question by exploiting reversible programming models and by developing techniques for fault ascription and explanation in component-based systems.

A common theme in this axis is the use and exploitation of causality information. Causality, *i.e.*, the logical dependence of an effect on a cause, has long been studied in disciplines such as philosophy [61], natural sciences, law [62], and statistics [63], but it has only recently emerged as an important focus of research in computer science. The analysis of logical causality has applications in many areas of computer science. For instance, tracking and analyzing logical causality between events in the execution of a concurrent system is required to ensure reversibility [58], to allow the diagnosis of faults in a complex concurrent system [54], or to enforce accountability [57], that is, designing systems in such a way that it can be determined without ambiguity whether a required safety or security property has been violated, and why. More generally, the goal of fault-tolerance can be understood as being to prevent certain causal chains from occurring by designing systems such that each causal chain either has its premises outside of the fault model (*e.g.*, by introducing redundancy [53]), or is broken (*e.g.*, by limiting fault propagation [65]).

SPECFUN Project-Team

3. Research Program

3.1. Studying special functions by computer algebra

Computer algebra manipulates symbolic representations of exact mathematical objects in a computer, in order to perform computations and operations like simplifying expressions and solving equations for "closed-form expressions". The manipulations are often fundamentally of algebraic nature, even when the ultimate goal is analytic. The issue of efficiency is a particular one in computer algebra, owing to the extreme swell of the intermediate values during calculations.

Our view on the domain is that research on the algorithmic manipulation of special functions is anchored between two paradigms:

- adopting linear differential equations as the right data structure for special functions,
- designing efficient algorithms in a complexity-driven way.

It aims at four kinds of algorithmic goals:

- algorithms combining functions,
- functional equations solving,
- multi-precision numerical evaluations,
- guessing heuristics.

This interacts with three domains of research:

- computer algebra, meant as the search for quasi-optimal algorithms for exact algebraic objects,
- symbolic analysis/algebraic analysis;
- experimental mathematics (combinatorics, mathematical physics, ...).

This view is made explicit in the present section.

3.1.1. Equations as a data structure

Numerous special functions satisfy linear differential and/or recurrence equations. Under a mild technical condition, the existence of such equations induces a finiteness property that makes the main properties of the functions decidable. We thus speak of *D-finite functions*. For example, 60 % of the chapters in the handbook [18] describe D-finite functions. In addition, the class is closed under a rich set of algebraic operations. This makes linear functional equations just the right data structure to encode and manipulate special functions. The power of this representation was observed in the early 1990s [70], leading to the design of many algorithms in computer algebra. Both on the theoretical and algorithmic sides, the study of D-finite functions shares much with neighbouring mathematical domains: differential algebra, D-module theory, differential Galois theory, as well as their counterparts for recurrence equations.

3.1.2. Algorithms combining functions

Differential/recurrence equations that define special functions can be recombined [70] to define: additions and products of special functions; compositions of special functions; integrals and sums involving special functions. Zeilberger's fast algorithm for obtaining recurrences satisfied by parametrised binomial sums was developed in the early 1990s already [71]. It is the basis of all modern definite summation and integration algorithms. The theory was made fully rigorous and algorithmic in later works, mostly by a group in RISC (Linz, Austria) and by members of the team [59], [67], [35], [33], [34], [54]. The past ÉPI Algorithms contributed several implementations (*gfun* [62], *Mgfun* [35]).

3.1.3. Solving functional equations

Encoding special functions as defining linear functional equations postpones some of the difficulty of the problems to a delayed solving of equations. But at the same time, solving (for special classes of functions) is a sub-task of many algorithms on special functions, especially so when solving in terms of polynomial or rational functions. A lot of work has been done in this direction in the 1990s; more intensively since the 2000s, solving differential and recurrence equations in terms of special functions has also been investigated.

3.1.4. Multi-precision numerical evaluation

A major conceptual and algorithmic difference exists for numerical calculations between data structures that fit on a machine word and data structures of arbitrary length, that is, *multi-precision* arithmetic. When multi-precision floating-point numbers became available, early works on the evaluation of special functions were just promising that "most" digits in the output were correct, and performed by heuristically increasing precision during intermediate calculations, without intended rigour. The original theory has evolved in a twofold way since the 1990s: by making computable all constants hidden in asymptotic approximations, it became possible to guarantee a *prescribed* absolute precision; by employing state-of-the-art algorithms on polynomials, matrices, etc, it became possible to have evaluation algorithms in a time complexity that is linear in the output size, with a constant that is not more than a few units. On the implementation side, several original works exist, one of which (*NumGfun* [58]) is used in our DDMF.

3.1.5. Guessing heuristics

"Differential approximation", or "Guessing", is an operation to get an ODE likely to be satisfied by a given approximate series expansion of an unknown function. This has been used at least since the 1970s and is a key stone in spectacular applications in experimental mathematics [32]. All this is based on subtle algorithms for Hermite–Padé approximants [22]. Moreover, guessing can at times be complemented by proven quantitative results that turn the heuristics into an algorithm [30]. This is a promising algorithmic approach that deserves more attention than it has received so far.

3.1.6. Complexity-driven design of algorithms

The main concern of computer algebra has long been to prove the feasibility of a given problem, that is, to show the existence of an algorithmic solution for it. However, with the advent of faster and faster computers, complexity results have ceased to be of theoretical interest only. Nowadays, a large track of works in computer algebra is interested in developing fast algorithms, with time complexity as close as possible to linear in their output size. After most of the more pervasive objects like integers, polynomials, and matrices have been endowed with fast algorithms for the main operations on them [41], the community, including ourselves, started to turn its attention to differential and recurrence objects in the 2000s. The subject is still not as developed as in the commutative case, and a major challenge remains to understand the combinatorics behind summation and integration. On the methodological side, several paradigms occur repeatedly in fast algorithms: "divide and conquer" to balance calculations, "evaluation and interpolation" to avoid intermediate swell of data, etc. [27].

3.2. Trusted computer-algebra calculations

3.2.1. Encyclopedias

Handbooks collecting mathematical properties aim at serving as reference, therefore trusted, documents. The decision of several authors or maintainers of such knowledge bases to move from paper books [18], [20], [63] to websites and wikis [0] allows for a more collaborative effort in proof reading. Another step toward further confidence is to manage to generate the content of an encyclopedia by computer-algebra programs, as is the case with the Wolfram Functions Site [0] or DDMF [0]. Yet, due to the lingering doubts about computer-algebra systems, some encyclopedias propose both cross-checking by different systems and handwritten companion paper proofs of their content [0]. As of today, there is no encyclopedia certified with formal proofs.

[0] for instance http://dlmf.nist.gov/ for special functions or http://oeis.org/ for integer sequences
[0] http://functions.wolfram.com/
[0] http://ddmf.msr-inria.inria.fr/1.9.1/ddmf

3.2.2. Computer algebra and symbolic logic

Several attempts have been made in order to extend existing computer-algebra systems with symbolic manipulations of logical formulas. Yet, these works are more about extending the expressivity of computer-algebra systems than about improving the standards of correctness and semantics of the systems. Conversely, several projects have addressed the communication of a proof system with a computer-algebra system, resulting in an increased automation available in the proof system, to the price of the uncertainty of the computations performed by this oracle.

3.2.3. Certifying systems for computer algebra

More ambitious projects have tried to design a new computer-algebra system providing an environment where the user could both program efficiently and elaborate formal and machine-checked proofs of correctness, by calling a general-purpose proof assistant like the Coq system. This approach requires a huge manpower and a daunting effort in order to re-implement a complete computer-algebra system, as well as the libraries of formal mathematics required by such formal proofs.

3.2.4. Semantics for computer algebra

The move to machine-checked proofs of the mathematical correctness of the output of computer-algebra implementations demands a prior clarification about the often implicit assumptions on which the presumably correctly implemented algorithms rely. Interestingly, this preliminary work, which could be considered as independent from a formal certification project, is seldom precise or even available in the literature.

3.2.5. Formal proofs for symbolic components of computer-algebra systems

A number of authors have investigated ways to organize the communication of a chosen computer-algebra system with a chosen proof assistant in order to certify specific components of the computer-algebra systems, experimenting various combinations of systems and various formats for mathematical exchanges. Another line of research consists in the implementation and certification of computer-algebra algorithms inside the logic [66], [46], [55] or as a proof-automation strategy. Normalization algorithms are of special interest when they allow to check results possibly obtained by an external computer-algebra oracle [38]. A discussion about the systematic separation of the search for a solution and the checking of the solution is already clearly outlined in [52].

3.2.6. Formal proofs for numerical components of computer-algebra systems

Significant progress has been made in the certification of numerical applications by formal proofs. Libraries formalizing and implementing floating-point arithmetic as well as large numbers and arbitrary-precision arithmetic are available. These libraries are used to certify floating-point programs, implementations of mathematical functions and for applications like hybrid systems.

3.3. Machine-checked proofs of formalized mathematics

To be checked by a machine, a proof needs to be expressed in a constrained, relatively simple formal language. Proof assistants provide facilities to write proofs in such languages. But, as merely writing, even in a formal language, does not constitute a formal proof just per se, proof assistants also provide a proof checker: a small and well-understood piece of software in charge of verifying the correctness of arbitrarily large proofs. The gap between the low-level formal language a machine can check and the sophistication of an average page of mathematics is conspicuous and unavoidable. Proof assistants try to bridge this gap by offering facilities, like notations or automation, to support convenient formalization methodologies. Indeed, many aspects, from the logical foundation to the user interface, play an important role in the feasibility of formalized mathematics inside a proof assistant.

[0] http://129.81.170.14/~vhm/Table.html

3.3.1. Logical foundations and proof assistants

While many logical foundations for mathematics have been proposed, studied, and implemented, type theory is the one that has been more successfully employed to formalize mathematics, to the notable exception of the Mizar system [56], which is based on set theory. In particular, the calculus of construction (CoC) [36] and its extension with inductive types (CIC) [37], have been studied for more than 20 years and been implemented by several independent tools (like Lego, Matita, and Agda). Its reference implementation, Coq [64], has been used for several large-scale formalizations projects (formal certification of a compiler back-end; four-color theorem). Improving the type theory underlying the Coq system remains an active area of research. Other systems based on different type theories do exist and, whilst being more oriented toward software verification, have been also used to verify results of mainstream mathematics (prime-number theorem; Kepler conjecture).

3.3.2. Computations in formal proofs

The most distinguishing feature of CoC is that computation is promoted to the status of rigorous logical argument. Moreover, in its extension CIC, we can recognize the key ingredients of a functional programming language like inductive types, pattern matching, and recursive functions. Indeed, one can program effectively inside tools based on CIC like Coq. This possibility has paved the way to many effective formalization techniques that were essential to the most impressive formalizations made in CIC.

Another milestone in the promotion of the computations-as-proofs feature of Coq has been the integration of compilation techniques in the system to speed up evaluation. Coq can now run realistic programs in the logic, and hence easily incorporates calculations into proofs that demand heavy computational steps.

Because of their different choice for the underlying logic, other proof assistants have to simulate computations outside the formal system, and indeed fewer attempts to formalize mathematical proofs involving heavy calculations have been made in these tools. The only notable exception, which was finished in 2014, the Kepler conjecture, required a significant work to optimize the rewriting engine that simulates evaluation in Isabelle/HOL.

3.3.3. Large-scale computations for proofs inside the Coq system

Programs run and proved correct inside the logic are especially useful for the conception of automated decision procedures. To this end, inductive types are used as an internal language for the description of mathematical objects by their syntax, thus enabling programs to reason and compute by case analysis and recursion on symbolic expressions.

The output of complex and optimized programs external to the proof assistant can also be stamped with a formal proof of correctness when their result is easier to *check* than to *find*. In that case one can benefit from their efficiency without compromising the level of confidence on their output at the price of writing and certify a checker inside the logic. This approach, which has been successfully used in various contexts, is very relevant to the present research project.

3.3.4. Relevant contributions from the Mathematical Component libraries

Representing abstract algebra in a proof assistant has been studied for long. The libraries developed by the MathComp project for the proof of the Odd Order Theorem provide a rather comprehensive hierarchy of structures; however, they originally feature a large number of instances of structures that they need to organize. On the methodological side, this hierarchy is an incarnation of an original work [40] based on various mechanisms, primarily type inference, typically employed in the area of programming languages. A large amount of information that is implicit in handwritten proofs, and that must become explicit at formalization time, can be systematically recovered following this methodology.

Small-scale reflection [43] is another methodology promoted by the MathComp project. Its ultimate goal is to ease formal proofs by systematically dealing with as many bureaucratic steps as possible, by automated computation. For instance, as opposed to the style advocated by Coq's standard library, decidable predicates are systematically represented using computable boolean functions: comparison on integers is expressed as

program, and to state that $a \leq b$ one compares the output of this program run on a and b with $true$. In many cases, for example when a and b are values, one can prove or disprove the inequality by pure computation.

The MathComp library was consistently designed after uniform principles of software engineering. These principles range from simple ones, like naming conventions, to more advanced ones, like generic programming, resulting in a robust and reusable collection of formal mathematical components. This large body of formalized mathematics covers a broad panel of algebraic theories, including of course advanced topics of finite group theory, but also linear algebra, commutative algebra, Galois theory, and representation theory. We refer the interested reader to the online documentation of these libraries [65], which represent about 150,000 lines of code and include roughly 4,000 definitions and 13,000 theorems.

Topics not addressed by these libraries and that might be relevant to the present project include real analysis and differential equations. The most advanced work of formalization on these domains is available in the HOL-Light system [48], [49], [50], although some existing developments of interest [25], [57] are also available for Coq. Another aspect of the MathComp libraries that needs improvement, owing to the size of the data we manipulate, is the connection with efficient data structures and implementations, which only starts to be explored.

3.3.5. User interaction with the proof assistant

The user of a proof assistant describes the proof he wants to formalize in the system using a textual language. Depending on the peculiarities of the formal system and the applicative domain, different proof languages have been developed. Some proof assistants promote the use of a declarative language, when the Coq and Matita systems are more oriented toward a procedural style.

The development of the large, consistent body of MathComp libraries has prompted the need to design an alternative and coherent language extension for the Coq proof assistant [45], [44], enforcing the robustness of proof scripts to the numerous changes induced by code refactoring and enhancing the support for the methodology of small-scale reflection.

The development of large libraries is quite a novelty for the Coq system. In particular any long-term development process requires the iteration of many refactoring steps and very little support is provided by most proof assistants, with the notable exception of Mizar [61]. For the Coq system, this is an active area of research.

STAMP Project-Team

3. Research Program

3.1. Theoretical background

The proof assistants that we consider provide both a programming language, where users can describe algorithms performing tasks in their domain of interest, and a logical language to reason about the programs, thus making it possible to ensure that the algorithms do solve the problems for which they were designed. trustability is gained because algorithms and logical statements provide multiple views of the same topic, thus making it possible to detect errors coming from mismatch between expected and established properties. The verification process is itself a logical process, where the computer can bring rigor in aligning expectations and guarantees.

The foundations of proof assistants rest on the very foundations of mathematics. As a consequence, all aspects of reasoning must be made completely explicit in the process of formally verifying an algorithm. All aspects of the formal verification of an algorithm are expressed in a discourse whose consistency is verified by the computer, so that unclear or intuitive arguments need to be replaced by precise logical inferences.

One of the foundational features on which we rely extensively is *Type Theory*. In this approach a very simple programming language is equiped with a powerful discipline to check the consistency of usage: types represent sets of data with similar behavior, functions represent algorithms mapping types to other types, and the consistency can be verified by a simple computer program, a *type-checker*. Although they can be verified by a simple program, types can express arbitrary complex objects or properties, so that the verification work lives in an interesting realm, where verifying proofs is decidable, but finding the proofs is undecidable.

This process for producing new algorithms and theorems is a novelty in the development of mathematical knowledge or algorithms, and new working methods must be devised for it to become a productive approach to high quality software development. Questions that arise are numerous. How do we avoid requiring human assistance to work on mundane aspects of proofs? How do we take advantage of all the progress made in automatic theorem proving? How do we organize the maintenance of ambitious corpora of formally verified knowledge in the long term?

To acquire hands-on expertise, we concentrate our activity on three aspects. The first one is foundational: we develop and maintain a library of mathematical facts that covers many aspects of algebra. In the past, we applied this library to proofs in group theory, but it is increasingly used for many different areas of mathematics and by other teams around the world, from combinatorics to elliptic cryptography, for instance. The second aspect is applicative: we develop a specific tool for proofs in cryptography, where we need to reason on the probability that opponents manage to access information we wish to protect. For this activity, we develop a specific proof system, relying on a wider set of automatic tools, with the objective of finding the tools that are well adapted to this domain and to attract users that are initially specialists in cryptography but not in formal verification. The third domain is robotics, as we believe that the current trend towards more an more autonomous robots and vehicles will raise questions of safety and trustability where formal verification can bring significant added value.

SUMO Project-Team

3. Research Program

3.1. Introduction

Since its creation in 2015, SUMO has successfully developed formal methods for large quantitative systems, in particular addressing verification, synthesis and control problems. Our current motivation is to expand this by putting emphasis on new concerns, such as algorithm efficiency, imprecision handling, and the more challenging objective of addressing incomplete or missing models. In the following we list a selection of detailed research goals, structured into four axes according to model classes: quantitative models, large systems, population models, and data-driven models. Some correspond to the pursuit of previously obtained results, others are more prospective.

3.2. Axis 1: Quantitative models

The analysis and control of quantitative models will remain at the heart of a large part of our research activities. In particular, we have two starting collaborative projects focusing on **timed models**, namely our ANR project TickTac and our collaboration with MERCE. The main expected outcome of TickTac is an open-source tool implementing the latest algorithms and allowing for quick prototyping of new algorithms. Several other topics will be explored in these collaborations, including robustness issues, game-theoretic problems, as well as the development of efficient algorithms, *e.g.* based on CEGAR approach or specifically designed for subclasses of automata (*e.g.* automata with few clocks and/or having a specific structure, as in [38]). Inspired by our collaboration with Alstom, we also aim at developing symbolic techniques for analysing non-linear timed models.

Stochastic models are another important focus for our research. On the one hand, we want to pursue our work on the optimization of non-standard properties for Markov decision processes, beyond the traditional verification questions, and explore *e.g.* long-run probabilities, and quantiles. Also, we aim at lifting our work on decisiveness from purely stochastic [36], [37] to non-deterministic and stochastic models in order to provide approximation schemes for the probability of (repeated) reachability properties in infinite-state Markov decision processes. On the other hand, in order to effectively handle large stochastic systems, we will pursue our work on approximation techniques. We aim at deriving simpler models, enjoying or preserving specific properties, and at determining the appropriate level of abstraction for a given system. One needs of course to quantify the approximation degrees (distances), and to preserve essential features of the original systems (explainability). This is a connection point between formal methods and the booming learning methods.

Regarding **diagnosis/opacity** issues, we will explore further the quantitative aspects. For diagnosis, the theory needs extensions to the case of incomplete or erroneous models, and to reconfigurable systems, in order to develop its applicability (see Sec. 3.6). There is also a need for non-binary causality analysis (*e.g.* performance degradations in complex systems). For opacity, we aim at quantifying the effort attackers must produce *vs* how much of a secret they can guess. We also plan to synthesize robust controllers resisting to sensor failures/attacks.

3.3. Axis 2: Large systems

Part of the background of SUMO is on the analysis and management of concurrent and modular/distributed systems, that we view as two main approaches to address state explosion problems. We will pursue the study of these models (including their quantitative features): verification of timed concurrent systems, robust distributed control of modular systems, resilient control to coalitions of attackers, distributed diagnosis, modular opacity analysis, distributed optimal planning, etc. Nevertheless, we have identified two new lines of effort, inspired by our application domains.

Reconfigurable systems. This is mostly motivated by applications at the convergence of virtualization techs with networking (Orange and Nokia PhDs). Software defined networks, either in the core (SDN/NFV) or at the edge (IoT) involve distributed systems that change structure constantly, to adapt to traffic, failures, maintenance, upgrades, etc. Traditional verification, control, diagnosis approaches (to mention only those) assume static and known models that can be handled as a whole. This is clearly insufficient here: one needs to adapt existing results to models that (sometimes automatically) change structure, incorporate new components/users or lose some, etc. At the same time, the programming paradigms for such systems (chaos monkey) incorporate resilience mechanisms, that should be considered by our models.

Hierarchical systems. Our experience with the regulation of subway lines (Alstom) revealed that large scale complex systems are usually described at a single level of granularity. Determining the appropriate granularity is a problem in itself. The control of such systems, with humans in the loop, can not be expressed at this single level, as tasks become too complex and require extremely skilled staff. It is rather desirable to describe models simultaneously at different levels of granularity, and to perform control at the appropriate level: humans in charge of managing the system by high level objectives, and computers in charge of implementing the appropriate micro-control sequences to achieve these tasks.

3.4. Axis 3: Population models

We want to step up our effort in parameterized verification of systems consisting of many identical components, so-called population models. In a nutshell our objectives summarize as "from Boolean to quantitative".

Inspired by our experience on the analysis of populations of yeasts, we aim at developping the quantitative analysis and control of population models, *e.g.* using Markov decision processes together with quantitative properties, and focusing on generating strategies with fast convergence.

As for broadcast networks, the challenge is to model the mobility of nodes (representing mobile ad hoc networks) in a faithful way. The obtained model should reflect on the one hand, the placement of nodes at a given time instant, and on the other hand, the physical movement of nodes over time. In this context, we will also use game theory techniques which allows one to study cooperative and conflictual behaviors of the nodes in the network, and to synthesize correct-by-design systems in adversarial environments.

As a new application area, we target randomized distributed algorithms. Our goal is to provide probabilistic variants of threshold automata [39] to represent fault-tolerant randomized distributed algorithms, designed for instance to solve the consensus problem. Most importantly, we then aim at developing new parameterized verification techniques, that will enable the automated verification of the correctness of such algorithms, as well as the assessment of their performances (in particular the expected time to termination).

In this axis, we will investigate whether fluid model checking and mean-field approximation techniques apply to our problems. More generally, we aim at a fruitful cross-fertilizing of these approaches with parameterized model-checking algorithms.

3.5. Axis 4: Data-driven models

In this axis, we will consider data-centric models, and in particular their application to crowd-sourcing. Many data-centric models such as Business Artifacts [40] orchestrate simple calls and answers to tasks performed by a single user. In a crowd-sourcing context, tasks are realized by pools of users, which may result in imprecise, uncertain and (partially) incompatible information. We thus need mechanisms to reconcile and fuse the various contributions in order to produce reliable information. Another aspect to consider concerns answers of higher-order: how to allow users to return intentional answers, under the form of a sub-workflow (coordinated set of tasks) which execution will provide the intended value. In the framework of the ANR Headwork we will build on formalisms such as GAG (guarded attribute grammars) or variants of business artifacts to propose formalisms adapted to crowd-sourcing applications, and tools to analyze them. To address imprecision, we will study techniques to handle fuzziness in user answers, will explore means to set incentives (rewards) dynamically, and to set competence requirements to guide the execution of a complex workflow, in order to achieve an objective with a desired level of quality.

In collaboration with Open Agora, CESPA and University of Yaoundé (Cameroun) we intend to implement in the GAG formalism some elements of argumentation theory (argumentation schemes, speech acts and dialogic games) in order to build a tool for the conduct of a critical discussion and the collaborative construction of expertise. The tool would incorporate point of view extraction (using clustering mechanisms), amendment management and consensus building mechanisms.

3.6. Transversal concern: missing models

We are concerned with one important lesson derived from our involvement in several application domains. Most of our background gets in force as soon as a perfect model of the system under study is available. Then verification, control, diagnosis, test, etc. can mobilize a solid background, or suggest new algorithmic problems to address. In numerous situations, however, assuming that a model is available is simply unrealistic. This is a major bottleneck for the impact of our research. We therefore intend to address this difficulty, in particular for the following domains.

- Model building for diagnosis. As a matter of fact, diagnosis theory hardly touches the ground to the extent that complete models of normal behavior are rarely available, and the identification of the appropriate abstraction level is unclear. Knowledge of faults and their effects is even less accessible. Also, the actual implemented systems may differ significantly from behaviors described in the norms. One therefore needs a theory for incomplete and erroneous models. Besides, one is often less bothered by partial observations than drowned by avalanches of alerts when malfunctions occur. Learning may come to the rescue, all the more that software systems may be deployed in sandpits and damaged for experimentation, thus allowing the collection of masses of labeled data. Competition on that theme clearly comes from Machine Learning techniques.
- Verification of large scale software. For some verification problems like the one we address in the IPL HAC-Specis, one does not have access to a formal model of the distributed program under study, but only to executions in a simulator. Formal verification poses new problems due to the difficulties to capture global states, to master state space explosion by gathering and exploiting concurrency information.
- Learning of stochastic models. Applications in bioinformatics often lead to large scale models, involving numerous chains of interactions between chemical species and/or cells. Fine grain models can be very precise, but very inefficient for inference or verification. Defining the appropriate levels of description/abstraction, given the available data and the verification goals, remains an open problem. This cannot be considered as a simple data fitting problem, as elements of biological knowledge must be combined with the data in order to preserve explainability of the phenomena.
- Testing and learning timed models: during conformance testing of a black-box implementation against its formal specification, one wants to detect non-conformances but may also want to learn the implementation model. Even though mixing testing and learning is not new, this is more recent and challenging for continuous-time models.
- Process mining. We intend to extend our work on process discovery using Petri net synthesis [35] by using negative information (*e.g.* execution traces identified as outliers) and quantitative information (probabilistic or fuzzy sets of execution traces) in order to infer more robust and precise models.

TAMIS Project-Team

3. Research Program

3.1. Axis 1: Vulnerability analysis

This axis proposes different techniques to discover vulnerabilities in systems. The outcomes of this axis are (a) new techniques to discover system vulnerabilities as well as to analyze them, and (b) to understand the importance of the hardware support.

Most existing approaches used at the engineering level rely on testing and fuzzing. Such techniques consist in simulating the system for various input values, and then checking that the result conforms to a given standard. The problem being the large set of inputs to be potentially tested. Existing solutions propose to extract significant sets by mutating a finite set of inputs. Other solutions, especially concolic testing developed at Microsoft, propose to exploit symbolic executions to extract constraints on new values. We build on those existing work, and extend them with recent techniques based on dissimilarity distances and learning. We also account for the execution environment, and study techniques based on the combination of timing attacks with fuzzing techniques to discover and classify classes of behavior of the system under test.

Techniques such as model checking and static analysis have been used for verifying several types of requirements such as safety and reliability. Recently, several works have attempted to adapt model checking to the detection of security issues. It has clearly been identified that this required to work at the level of binary code. Applying formal techniques to such code requires the development of disassembly techniques to obtain a semantically well-defined model. One of the biggest issues faced with formal analysis is the state space explosion problem. This problem is amplified in our context as representations of data (such as stack content) definitively blow up the state space. We propose to use statistical model checking (SMC) of rare events to efficiently identify problematic behaviors.

We also seek to understand vulnerabilities at the architecture and hardware levels. Particularly, we evaluate vulnerabilities of the interfaces and how an adversary could use them to get access to core assets in the system. One particular mechanism to be investigated is the DMA and the so-called Trustzone. An ad-hoc technique to defend against adversarial DMA-access to memory is to keep key material exclusively in registers. This implies co-analyzing machine code and an accurate hardware model.

3.2. Axis 2: Malware analysis

Axis 1 is concerned with vulnerabilities. Such vulnerabilities can be exploited by an attacker in order to introduce malicious behaviors in a system. Another method to identify vulnerabilities is to analyze malware that exploits them. However, modern malware has a wide variety of analysis avoidance techniques. In particular, attackers obfuscate the code leading to a security exploit. For doing so, recent black hat research suggests hiding constants in program choices via polynomials. Such techniques hinder forensic analysis by making detailed analysis labor intensive and time consuming. The objective of research axis 2 is to obtain a full tool chain for malware analysis starting from (a) the observability of the malware via deobfuscation, and (b) the analysis of the resulting binary file. A complementary objective is to understand how hardware attacks can be exploited by malwares.

We first investigate obfuscation techniques. Several solutions exist to mitigate the packer problem. As an example, we try to reverse the packer and remove the environment evaluation in such a way that it performs the same actions and outputs the resulting binary for further analysis. There is a wide range of techniques to obfuscate malware, which includes flattening and virtualization. We will produce a taxonomy of both techniques and tools. We will first give a particular focus to control flow obfuscation via mixed Boolean algebra, which is highly deployed for malware obfuscation. We recently showed that a subset of them can be broken via SAT-solving and synthesis. Then, we will expand our research to other obfuscation techniques.

Once the malware code has been unpacked/deobfuscated, the resulting binary still needs to be fully understood. Advanced malware often contains multiple stages, multiple exploits and may unpack additional features based on its environment. Ensuring that one understands all interesting execution paths of a malware sample is related to enumerating all of the possible execution paths when checking a system for vulnerabilities. The main difference is that in one case we are interested in finding vulnerabilities and in the other in finding exploitative behavior that may mutate. Still, some of the techniques of Axis 1 can be helpful in analyzing malware. The main challenge for axis 2 is thus to adapt the tools and techniques to deal with binary programs as inputs, as well as the logic used to specify malware behavior, including behavior with potentially rare occurrences. Another challenge is to take mutation into account, which we plan to do by exploiting mining algorithms.

Most recent attacks against hardware are based on fault injection which dynamically modifies the semantics of the code. We demonstrated the possibility to obfuscate code using constraint solver in such a way that the code becomes intentionally hostile while hit by a laser beam. This new form of obfuscation opens a new challenge for secure devices where malicious programs can be designed and uploaded that defeat comprehensive static analysis tools or code reviews, due to their multi-semantic nature. We have shown on several products that such an attack cannot be mitigated with the current defenses embedded in Java cards. In this research, we first aim at extending the work on fault injection, then at developing new techniques to analyze such hostile code. This is done by proposing formal models of fault injection, and then reusing results from our work on obfuscation/deobfuscation.

TEA Project-Team

3. Research Program

3.1. Previous Works

The challenges of team TEA support the claim that sound Cyber-Physical System design (including embedded, reactive, and concurrent systems altogether) should consider multi-form time models as a central aspect. In this aim, architectural specifications found in software engineering are a natural focal point to start from. Architecture descriptions organize a system model into manageable components, establish clear interfaces between them, collect domain-specific constraints and properties to help correct integration of components during system design. The definition of a formal design methodology to support heterogeneous or multi-form models of time in architecture descriptions demands the elaboration of sound mathematical foundations and the development of formal calculi and methods to instrument them.

System design based on the “synchronous paradigm” has focused the attention of many academic and industrial actors on abstracting non-functional implementation details from system design. This elegant design abstraction focuses on the logic of interaction in reactive programs rather than their timed behavior, allowing to secure functional correctness while remaining an intuitive programming model for embedded systems. Yet, it corresponds to embedded technologies of single cores and synchronous buses from the 90s, and may hardly cover the semantic diversity of distribution, parallelism, heterogeneity, of cyber-physical systems found in 21st century Internet-connected, true-time^{TM}-synchronized clouds, of tomorrow’s grids.

By contrast with a synchronous hypothesis, yet from the same era, the polychronous MoCC is inherently capable of describing multi-clock abstractions of GALS systems. Polychrony is implemented in the data-flow specification language Signal, available in the Eclipse project POP [0] and in the CCSL standard [0] available from the TimeSquare project. Both provide tooled infrastructures to refine high-level specifications into real-time streaming applications or locally synchronous and globally asynchronous systems, through a series of model analysis, verification, and synthesis services. These tool-supported refinement and transformation techniques can assist the system engineer from the earliest design stages of requirement specification to the latest stages of synthesis, scheduling and deployment. These characteristics make polychrony much closer to the required semantic for compositional, refinement-based, architecture-driven, system design.

While polychrony was a step ahead of the traditional synchronous hypothesis, CCSL is a leap forward from synchrony and polychrony. The essence of CCSL is “multi-form time” toward addressing all of the domain-specific physical, electronic and logical aspects of cyber-physical system design.

3.2. Timed Modeling

To formalize timed semantics for system design, we shall rely on algebraic representations of time as clocks found in previous works and introduce a paradigm of "time system" (types that represent time) in a way reminiscent to CCSL. Just as a type system abstracts data carried along operations in a program, a time system abstracts the causal interaction of that program module or hardware element with its environment, its pre and post conditions, its assumptions and guarantees, either logical or numerical, discrete or continuous. Some fundamental concepts of the time systems we envision are present in the clock calculi found in data-flow synchronous languages like Signal or Lustre, yet bound to a particular model of timed concurrency.

[0] *Polychrony on Polarsys*, https://www.polarsys.org/projects/polarsys.pop
[0] *Clock Constraints in UML/MARTE CCSL*. C. André, F. Mallet. RR-6540. Inria, 2008. http://hal.inria.fr/inria-00280941

In particular, the principle of refinement type systems [0], is to associate information (data-types) inferred from programs and models with properties pertaining, for instance, to the algebraic domain on their value, or any algebraic property related to its computation: effect, memory usage, pre-post condition, value-range, cost, speed, time, temporal logic [0]. Being grounded on type and domain theories, a time system should naturally be equipped with program analysis techniques based on type inference (for data-type inference) or abstract interpretation (for program properties inference) to help establish formal relations between heterogeneous component "types". Just as a time calculus may formally abstract timed concurrent behaviors of system components, timed relations (abstraction and refinement) represent interaction among components.

Scalability requires the use of assume-guarantee reasoning to allow modularity and to facilitate composition by behavioral sub-typing, in the spirit of the (static) contract-based formalism proposed by Passerone et al. [0]. Verification problems encompassing heterogeneously timed specifications are common and of great variety: checking correctness between abstract (e.g. the synchronous hypothesis) and concrete time models (e.g. real-time architectures) relates to desynchronisation (from synchrony to asynchrony) and scheduling analysis (from synchronous data-flow to hardware). More generally, they can be perceived from heterogeneous timing viewpoints (e.g. mapping a synchronous-time software on a real-time middle-ware or hardware).

This perspective demands capabilities to use abstraction and refinement mechanisms for time models (using simulation, refinement, bi-simulation, equivalence relations) but also to prove more specific properties (synchronization, determinism, endochrony). All this formalization effort will allow to effectively perform the tooled validation of common cross-domain properties (e.g. cost v.s. power v.s. performance v.s. software mapping) and tackle problems such as these integrating constraints of battery capacity, on-board CPU performance, available memory resources, software schedulability, to logical software correctness and plant controllability.

3.3. Modeling Architectures

To address the formalization of such cross-domain case studies, modeling the architecture formally plays an essential role. An architectural model represents components in a distributed system as boxes with well-defined interfaces, connections between ports on component interfaces, and specifies component properties that can be used in analytical reasoning about the model. Several architectural modeling languages for embedded systems have emerged in recent years, including the SAE AADL [0], SysML [0], UML MARTE [0].

In system design, an architectural specification serves several important purposes. First, it breaks down a system model into components of manageable size and complexity, to establish clear interfaces between components. In this way, complexity becomes manageable by hiding details that are not relevant at a given level of abstraction. Clear, formally defined, component interfaces allow us to avoid integration problems at the implementation phase. Connections between components, which specify how components interact with each other, help propagate the effects of a change in one component to the linked components.

Most importantly, an architectural model is a repository to share knowledge about the system being designed. This knowledge can be represented as requirements, design artifacts, component implementations, held together by a structural backbone. Such a repository enables automatic generation of analytical models for different aspects of the system, such as timing, reliability, security, performance, energy, etc. Since all the models are generated from the same source, the consistency of assumptions w.r.t. guarantees, of abstractions w.r.t. refinements, used for different analyses becomes easier, and can be properly ensured in a design methodology based on formal verification and synthesis methods.

Related works in this aim, and closer in spirit to our approach (to focus on modeling time) are domain-specific languages such as Prelude [0] to model the real-time characteristics of embedded software architectures.

[0] *Abstract Refinement Types*. N. Vazou, P. Rondon, and R. Jhala. European Symposium on Programming. Springer, 2013.
[0] *LTL types FRP*. A. Jeffrey. Programming Languages meets Program Verification.
[0] *A contract-based formalism for the specification of heterogeneous systems*. L. Benvenistu, et al. FDL, 2008
[0] *Architecture Analysis and Design Language*, AS-5506. SAE, 2004. http://standards.sae.org/as5506b
[0] *System modeling Language*. OMG, 2007. http://www.omg.org/spec/SysML
[0] *UML Profile for MARTE*. OMG, 2009. http://www.omg.org/spec/MARTE
[0] *The Prelude language*. LIFL and ONERA, 2012. http://www.lifl.fr/~forget/prelude.html

Conversely, standard architecture description languages could be based on algebraic modeling tools, such as interface theories with the ECDAR tool [0].

In project TEA, it takes form by the normalization of the AADL standard's formal semantics and the proposal of a time specification annex in the form of related standards, such as CCSL, to model concurrency, time and physical properties, and PSL, to model timed traces.

3.4. Scheduling Theory

Based on sound formalization of time and CPS architectures, real-time scheduling theory provides tools for predicting the timing behavior of a CPS which consists of many interacting software and hardware components. Expressing parallelism among software components is a crucial aspect of the design process of a CPS. It allows for efficient partition and exploitation of available resources.

The literature about real-time scheduling [0] provides very mature schedulability tests regarding many scheduling strategies, preemptive or non-preemptive scheduling, uniprocessor or multiprocessor scheduling, etc. Scheduling of data-flow graphs has also been extensively studied in the past decades.

A milestone in this prospect is the development of abstract affine scheduling techniques [0]. It consists, first, of approximating task communication patterns (e.g. between Safety-Critical Java threads) using cyclo-static data-flow graphs and affine functions. Then, it uses state of the art ILP techniques to find optimal schedules and to concretize them as real-time schedules in the program implementations [00].

Abstract scheduling, or the use of abstraction and refinement techniques in scheduling borrowed to the theory of abstract interpretation [0] is a promising development toward tooled methodologies to orchestrate thousands of heterogeneous hardware/software blocks on modern CPS architectures (just consider modern cars or aircrafts). It is an issue that simply defies the state of the art and known bounds of complexity theory in the field, and consequently requires a particular focus.

To develop the underlying theory of this promising research topic, we first need to deepen the theoretical foundation to establish links between scheduling analysis and abstract interpretation. A theory of time systems would offer the ideal framework to pursue this development. It amounts to representing scheduling constraints, inferred from programs, as types or contract properties. It allows to formalize the target time model of the scheduler (the architecture, its middle-ware, its real-time system) and defines the basic concepts to verify assumptions made in one with promises offered by the other: contract verification or, in this case, synthesis.

3.5. Verified programming for system design

The IoT is a network of devices that sense, actuate and change our immediate environment. Against this fundamental role of sensing and actuation, design of edge devices often considers actions and event timings to be primarily software implementation issues: programming models for IoT abstract even the most rudimentary information regarding timing, sensing and the effects of actuation. As a result, applications programming interfaces (API) for IoT allow wiring systems fast without any meaningful assertions about correctness, reliability or resilience.

We make the case that the "API glue" must give way to a logical interface expressed using contracts or refinement types. Interfaces can be governed by a calculus – a refinement type calculus – to enable reasoning on time, sensing and actuation, in a way that provides both deep specification refinement, for mechanized verification of requirements, and multi-layered abstraction, to support compositionality and scalability, from one end of the system to the other.

[0] *PyECDAR, timed games for timed specifications*. Inria, 2013. https://project.inria.fr/pyecdar
[0] *A survey of hard real-time scheduling for multiprocessor systems*. R. I. Davis and A. Burns. *ACM Computing Survey* 43(4), 2011.
[0] *Buffer minimization in EDF scheduling of data-flow graphs*. A. Bouakaz and J.-P. Talpin. LCTES, ACM, 2013.
[0] *ADFG for the synthesis of hard real-time applications*. A. Bouakaz, J.-P. Talpin, J. Vitek. ACSD, IEEE, June 2012.
[0] *Design of SCJ Level 1 Applications Using Affine Abstract Clocks*. A. Bouakaz and J.-P. Talpin. SCOPES, ACM, 2013.
[0] *La vérification de programmes par interprétation abstraite*. P. Cousot. Séminaire au Collège de France, 2008.

Our project seeks to elevate the "function as type" paradigm to that of "system as type": to define a refinement type calculus based on concepts of contracts for reasoning on networked devices and integrate them as cyber-physical systems [0]. An invited paper [0] outlines our progress with respect to this aim and plans towards building a verified programming environment for networked IoT devices: we propose a type-driven approach to verifying and building safe and secure IoT applications.

Accounting for such constrains in a more principled fashion demands reasoning about the composition of all the software and hardware components of the application. Our proposed framework takes a step in this direction by (1) using refinement types to make make physical constraints explicit and (2) imposing an event-driven programming discipline to simplify the reasoning of system-wide properties to that of an event queue. In taking this approach, our approach would make it possible for a developer to build a verified IoT application by ensuring that a well-typed program cannot violate the physical constraints of its architecture and environment.

[0] Refinement types for system design. Jean-Pierre Talpin. FDL'18 keynote.

[0] Steps toward verified programming of embedded computing systems. Jean-Pierre Talpin, Jean-Joseph Marty, Deian Stefan, Shravan Nagarayan, Rajesh Gupta, DATE'18.

TOCCATA Project-Team

3. Research Program

3.1. Research Program

3.1.1. Panorama of Deductive Verification

There are two main families of approaches for deductive verification. Methods in the first family build on top of mathematical proof assistants (e.g., Coq, Isabelle) in which both the model and the program are encoded; the proof that the program meets its specification is typically conducted in an interactive way using the underlying proof construction engine. Methods from the second family proceed by the design of standalone tools taking as input a program in a particular programming language (e.g., C, Java) specified with a dedicated annotation language (e.g., ACSL [49], JML [56]) and automatically producing a set of mathematical formulas (the *verification conditions*) which are typically proved using automatic provers (e.g., Z3 [70], Alt-Ergo [58], CVC4 [48]).

The first family of approaches usually offers a higher level of assurance than the second, but also demands more work to perform the proofs (because of their interactive nature) and makes them less easy to adopt by industry. Moreover, they generally do not allow to directly analyze a program written in a mainstream programming language like Java or C. The second kind of approaches has benefited in the past years from the tremendous progress made in SAT and SMT solving techniques, allowing more impact on industrial practices, but suffers from a lower level of trust: in all parts of the proof chain (the model of the input programming language, the VC generator, the back-end automatic prover), potential errors may appear, compromising the guarantee offered. Moreover, while these approaches are applied to mainstream languages, they usually support only a subset of their features.

3.1.2. Overall Goals of the Toccata Project

One of our original skills is the ability to conduct proofs by using automatic provers and proof assistants at the same time, depending on the difficulty of the program, and specifically the difficulty of each particular verification condition. We thus believe that we are in a good position to propose a bridge between the two families of approaches of deductive verification presented above. Establishing this bridge is one of the goals of the Toccata project: we want to provide methods and tools for deductive program verification that can offer both a high amount of proof automation and a high guarantee of validity. Indeed, an axis of research of Toccata is the development of languages, methods and tools that are themselves formally proved correct.

In industrial applications, numerical calculations are very common (e.g. control software in transportation). Typically they involve floating-point numbers. Some of the members of Toccata have an internationally recognized expertise on deductive program verification involving floating-point computations. Our past work includes a new approach for proving behavioral properties of numerical C programs using Frama-C/Jessie [47], various examples of applications of that approach [54], the use of the Gappa solver for proving numerical algorithms [68], an approach to take architectures and compilers into account when dealing with floating-point programs [55], [66]. We also contributed to the Handbook of Floating-Point Arithmetic [65]. A representative case study is the analysis and the proof of both the method error and the rounding error of a numerical analysis program solving the one-dimension acoustic wave equation [52] [51]. Our experience led us to a conclusion that verification of numerical programs can benefit a lot from combining automatic and interactive theorem proving [53], [54], [59]. Verification of numerical programs is another main axis of Toccata.

Our scientific programme detailed below is structured into four axes:

1. Foundations and spreading of deductive program verification;
2. Reasoning on mutable memory in program verification;
3. Verification of Computer Arithmetic;
4. Spreading Formal Proofs.

Let us conclude with more general considerations about our agenda of the next four years: we want to keep on

- with general audience actions;
- industrial transfer, in particular through an extension of the perimeter of the ProofInUse joint lab.

3.2. Foundations and spreading of deductive program verification

Permanent researchers: S. Conchon, J.-C. Filliâtre, C. Marché, G. Melquiond, A. Paskevich

This axis covers the central theme of the team: deductive verification, from the point of view of its foundations but also our will to spread its use in software development. The general motto we want to defend is "deductive verification for the masses". A non-exhaustive list of subjects we want to address is as follows.

- The verification of general-purpose algorithms and data structures: the challenge is to discover adequate invariants to obtain a proof, in the most automatic way as possible, in the continuation of the current VOCaL project and the various case studies presented in Axis 4 below.
- Uniform approaches to obtain correct-by-construction programs and libraries, in particular by automatic extraction of executable code (in OCaml, C, CakeML, etc.) from verified programs, and including innovative general methods like advanced ghost code, ghost monitoring, etc.
- Automated reasoning dedicated to deductive verification, so as to improve proof automation; improved combination of interactive provers and fully automated ones, proof by reflection.
- Improved feedback in case of proof failures: based on generation of counterexamples, or symbolic execution, or possibly randomized techniques à la quickcheck.
- Reduction of the trusted computing base in our toolchains: production of certificates from automatic proofs, for goal transformations (like those done by Why3), and from the generation of VCs

A significant part of the work achieved in this axis is related to the Why3 toolbox and its ecosystem, displayed on Figure 1 . The boxes in red background correspond to the tools we develop in the Toccata team.

3.3. Reasoning on mutable memory in program verification

Permanent researchers: J.-C. Filliâtre, C. Marché, G. Melquiond, A. Paskevich

This axis concerns specifically the techniques for reasoning on programs where aliasing is the central issue. It covers the methods based on type-based alias analysis and related memory models, on specific program logics such as separation logics, and extended model-checking. It concerns the application on analysis of C or C++ codes, on Ada codes involving pointers, but also concurrent programs in general. The main topics planned are:

- The study of advanced type systems dedicated to verification, for controlling aliasing, and their use for obtaining easier-to-prove verification conditions. Modern typing system in the style of Rust, involving ownership and borrowing, will be considered.
- The design of front-ends of Why3 for the proofs of programs where aliasing cannot be fully controlled statically, via adequate memory models, aiming in particular at extraction to C; and also for concurrent programs.
- The continuation of fruitful work on concurrent parameterized systems, and its corresponding specific SMT-based model-checking.
- Concurrent programming on weak memory models, on one hand as an extension of parameterized systems above, but also in the specific context of OCaml multicore (http://ocamllabs.io/doc/multicore.html).
- In particular in the context of the ProofInUse joint lab, design methods for Ada, C, C++ or Java using memory models involving fine-grain analysis of pointers. Rust programs could be considered as well.

3.4. Verification of Computer Arithmetic

Permanent researchers: S. Boldo, C. Marché, G. Melquiond

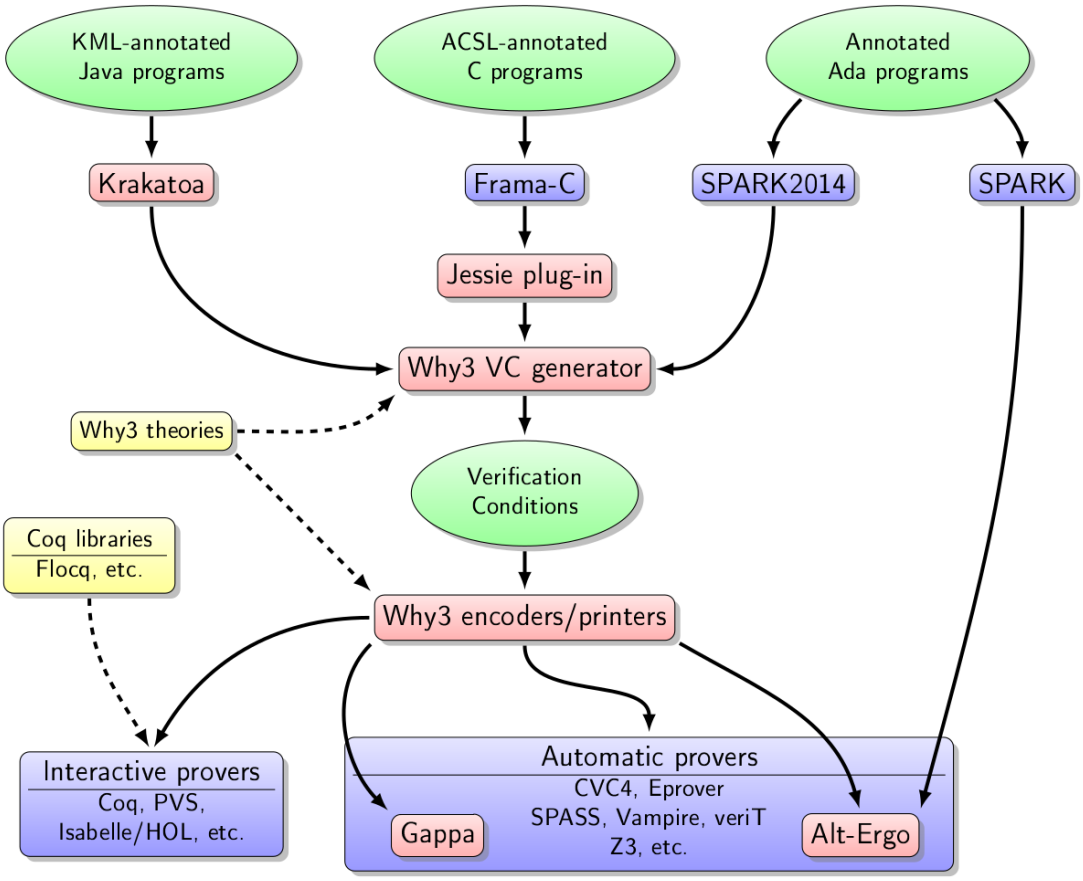


Figure 1. The Why3 ecosystem

We of course want to keep this axis which is a major originality of Toccata. The main topics of the next 4 years will be:

- Fundamental studies concerning formalization of floating-point computations, algorithms, and error analysis. Related to numerical integration, we will develop the relationships between mathematical stability and floating-point stability of numerical schemes.
- A significant effort dedicated to verification of numerical programs written in C, Ada, C++. This involves combining specifications in real numbers and computation in floating-point, and underlying automated reasoning techniques with floating-point numbers and real numbers. A new approach we have in mind concerns some variant of symbolic execution of both code and specifications involving real numbers.
- We have not yet studied embedded systems. Our approach is first to tackle numerical filters. This requires more results on fixed-point arithmetic and a careful study of overflows.
- Also a specific focus on arbitrary precision integer arithmetic, in the continuation of the ongoing PhD thesis of R. Rieu-Helft.

3.5. Spreading Formal Proofs

Permanent researchers: S. Boldo, S. Conchon, J.-C. Filliâtre, C. Marché, G. Melquiond, A. Paskevich

This axis covers applications in general. The applications we currently have in mind are:

- Hybrid Systems, i.e., systems mixing discrete and continuous transitions. This theme covers many aspects such as general techniques for formally reasoning of differential equations, and extending SMT-based reasoning. The challenge is to support both abstract mathematical reasoning and concrete program execution (e.g., using floating-point representation). Hybrid systems will be a common effort with other members of the future laboratory joint with LSV of ENS Cachan.
- Applied mathematics, in the continuation of the current efforts towards verification of Finite Element Method. It has only been studied in the mathematical point of view during this period. We plan to also consider their floating-point behavior and a demanding application is that of molecular simulation exhibited in the new EMC2 project. The challenge here is both in the mathematics to be formalized, in the numerical errors that have never been studied (and that may be huge in specific cases), and in the size of the programs, which requires that our tools scale.
- Continuation of our work on analysis of shell scripts. The challenge is to be able to analyze a large number of scripts (more than 30,000 in the corpus of Debian packages installation scripts) in an automatic manner. An approach that will be considered is some form of symbolic execution.
- Explore proof tools for mathematics, in particular automated reasoning for real analysis (application: formalization of the weak Goldbach conjecture), and in number theory.
- Obtain and distribute verified OCaml libraries, as expected outcome of the VOCaL project.
- Formalization of abstract interpretation and WP calculi: in the continuation of the former project Verasco, and an ongoing project proposal joint with CEA List. The difficulty of achieving full verification of such tools will be mitigated by use of certificate techniques.

VERIDIS Project-Team

3. Research Program

3.1. Automated and Interactive Theorem Proving

The VeriDis team gathers experts in techniques and tools for automatic deduction and interactive theorem proving, and specialists in methods and formalisms designed for the development of trustworthy concurrent and distributed systems and algorithms. Our common objective is twofold: first, we wish to advance the state of the art in automated and interactive theorem proving, and their combinations. Second, we work on making the resulting technology available for the computer-aided verification of distributed systems and protocols. In particular, our techniques and tools are intended to support sound methods for the development of trustworthy distributed systems that scale to algorithms relevant for practical applications.

VeriDis members from Saarbrücken are developing the SPASS [10] workbench. It currently consists of one of the leading automated theorem provers for first-order logic based on the superposition calculus [56] and a theory solver for linear arithmetic.

In a complementary approach to automated deduction, VeriDis members from Nancy work on techniques for integrating reasoners for specific theories. They develop veriT [1], an SMT [0] solver that combines decision procedures for different fragments of first-order logic. The veriT solver is designed to produce detailed proofs; this makes it particularly suitable as a component of a robust cooperation of deduction tools.

Finally, VeriDis members design effective quantifier elimination methods and decision procedures for algebraic theories, supported by their efficient implementation in the Redlog system [4].

An important objective of this line of work is the integration of theories in automated deduction. Typical theories of interest, including fragments of arithmetic, are difficult or impossible to express in first-order logic. We therefore explore efficient, modular techniques for integrating semantic and syntactic reasoning methods, develop novel combination results and techniques for quantifier instantiation. These problems are addressed from both sides, i.e. by embedding decision procedures into the superposition framework or by allowing an SMT solver to accept axiomatizations for plug-in theories. We also develop specific decision procedures for theories such as non-linear real arithmetic that are important when reasoning about certain classes of (e.g., real-time) systems but that also have interesting applications beyond verification.

We rely on interactive theorem provers for reasoning about specifications at a high level of abstraction when fully automatic verification is not (yet) feasible. An interactive proof platform should help verification engineers lay out the proof structure at a sufficiently high level of abstraction; powerful automatic plug-ins should then discharge the resulting proof steps. Members of VeriDis have ample experience in the specification and subsequent machine-assisted, interactive verification of algorithms. In particular, we participate in a project at the joint Microsoft Research-Inria Centre on the development of methods and tools for the formal proof of TLA$^+$ [66] specifications. Our prover relies on a declarative proof language, and calls upon several automatic backends [3]. Trust in the correctness of the overall proof can be ensured when the backends provide justifications that can be checked by the trusted kernel of a proof assistant. During the development of a proof, most obligations that are passed to the prover actually fail – for example, because necessary information is not present in the context or because the invariant is too weak, and we are interested in explaining failed proof attempts to the user, in particular through the construction of counter-models.

[0] Satisfiability Modulo Theories [58]

3.2. Formal Methods for Developing and Analyzing Algorithms and Systems

Theorem provers are not used in isolation, but they support the application of sound methodologies for modeling and verifying systems. In this respect, members of VeriDis have gained expertise and recognition in making contributions to formal methods for concurrent and distributed algorithms and systems [2], [9], and in applying them to concrete use cases. In particular, the concept of *refinement* [55], [57], [70] in state-based modeling formalisms is central to our approach because it allows us to present a rational (re)construction of system development. An important goal in designing such methods is to establish precise proof obligations, many of which can be discharged by automatic tools. This requires taking into account specific characteristics of certain classes of systems and tailoring the model to concrete computational models. Our research in this area is supported by carrying out case studies for academic and industrial developments. This activity benefits from and influences the development of our proof tools.

In this line of work, we investigate specific development and verification patterns for particular classes of algorithms, in order to reduce the work associated with their verification. We are also interested in applications of formal methods and their associated tools to the development of systems that underlie specific certification requirements in the sense of, e.g., Common Criteria. Finally, we are interested in the adaptation of model checking techniques for verifying actual distributed programs, rather than high-level models.

Today, the formal verification of a new algorithm is typically the subject of a PhD thesis, if it is addressed at all. This situation is not sustainable given the move towards more and more parallelism in mainstream systems: algorithm developers and system designers must be able to productively use verification tools for validating their algorithms and implementations. On a high level, the goal of VeriDis is to make formal verification standard practice for the development of distributed algorithms and systems, just as symbolic model checking has become commonplace in the development of embedded systems and as security analysis for cryptographic protocols is becoming standard practice today. Although the fundamental problems in distributed programming are well-known, they pose new challenges in the context of modern system paradigms, including ad-hoc and overlay networks or peer-to-peer systems, and they must be integrated for concrete applications.

ACUMES Project-Team

3. Research Program

3.1. Research directions

The project develops along the following two axes:

- modeling complex systems through novel (unconventional) PDE systems, accounting for multi-scale phenomena and uncertainty;
- optimization and optimal control algorithms for systems governed by the above PDE systems.

These themes are motivated by the specific problems treated in the applications, and represent important and up-to-date issues in engineering sciences. For example, improving the design of transportation means and civil buildings, and the control of traffic flows, would result not only in better performances of the object of the optimization strategy (vehicles, buildings or road networks level of service), but also in enhanced safety and lower energy consumption, contributing to reduce costs and pollutant emissions.

3.1.1. PDE models accounting for multi-scale phenomena and uncertainties

Dynamical models consisting of evolutionary PDEs, mainly of hyperbolic type, appear classically in the applications studied by the previous Project-Team Opale (compressible flows, traffic, cell-dynamics, medicine, etc). Yet, the classical purely macroscopic approach is not able to account for some particular phenomena related to specific interactions occurring at smaller scales. These phenomena can be of greater importance when dealing with particular applications, where the "first order" approximation given by the purely macroscopic approach reveals to be inadequate. We refer for example to self-organizing phenomena observed in pedestrian flows [126], or to the dynamics of turbulent flows for which large scale / small scale vortical structures interfere [155].

Nevertheless, macroscopic models offer well known advantages, namely a sound analytical framework, fast numerical schemes, the presence of a low number of parameters to be calibrated, and efficient optimization procedures. Therefore, we are convinced of the interest of keeping this point of view as dominant, while completing the models with information on the dynamics at the small scale / microscopic level. This can be achieved through several techniques, like hybrid models, homogenization, mean field games. In this project, we will focus on the aspects detailed below.

The development of adapted and efficient numerical schemes is a mandatory completion, and sometimes ingredient, of all the approaches listed below. The numerical schemes developed by the team are based on finite volumes or finite elements techniques, and constitute an important tool in the study of the considered models, providing a necessary step towards the design and implementation of the corresponding optimization algorithms, see Section 3.1.2 .

3.1.1.1. Micro-macro couplings

Modeling of complex problems with a dominant macroscopic point of view often requires couplings with small scale descriptions. Accounting for systems heterogeneity or different degrees of accuracy usually leads to coupled PDE-ODE systems.

In the case of heterogeneous problems the coupling is "intrinsic", i.e. the two models evolve together and mutually affect each-other. For example, accounting for the impact of a large and slow vehicle (like a bus or a truck) on traffic flow leads to a strongly coupled system consisting of a (system of) conservation law(s) coupled with an ODE describing the bus trajectory, which acts as a moving bottleneck.The coupling is realized through a local unilateral moving constraint on the flow at the bus location, see [94] for an existence result and [77], [93] for numerical schemes.

If the coupling is intended to offer higher degree of accuracy at some locations, a macroscopic and a microscopic model are connected through an artificial boundary, and exchange information across it through suitable boundary conditions. See [84], [115] for some applications in traffic flow modelling, and [106], [111], [113] for applications to cell dynamics.

The corresponding numerical schemes are usually based on classical finite volume or finite element methods for the PDE, and Euler or Runge-Kutta schemes for the ODE, coupled in order to take into account the interaction fronts. In particular, the dynamics of the coupling boundaries require an accurate handling capturing the possible presence of non-classical shocks and preventing diffusion, which could produce wrong solutions, see for example [77], [93].

We plan to pursue our activity in this framework, also extending the above mentioned approaches to problems in two or higher space dimensions, to cover applications to crowd dynamics or fluid-structure interaction.

3.1.1.2. Micro-macro limits

Rigorous derivation of macroscopic models from microscopic ones offers a sound basis for the proposed modeling approach, and can provide alternative numerical schemes, see for example [85], [96] for the derivation of Lighthill-Whitham-Richards [138], [154] traffic flow model from Follow-the-Leader and [107] for results on crowd motion models (see also [128]). To tackle this aspect, we will rely mainly on two (interconnected) concepts: measure-valued solutions and mean-field limits.

The notion of **measure-valued solutions** for conservation laws was first introduced by DiPerna [97], and extensively used since then to prove convergence of approximate solutions and deduce existence results, see for example [108] and references therein. Measure-valued functions have been recently advocated as the appropriate notion of solution to tackle problems for which analytical results (such as existence and uniqueness of weak solutions in distributional sense) and numerical convergence are missing [64], [110]. We refer, for example, to the notion of solution for non-hyperbolic systems [116], for which no general theoretical result is available at present, and to the convergence of finite volume schemes for systems of hyperbolic conservation laws in several space dimensions, see [110].

In this framework, we plan to investigate and make use of measure-based PDE models for vehicular and pedestrian traffic flows. Indeed, a modeling approach based on (multi-scale) time-evolving measures (expressing the agents probability distribution in space) has been recently introduced (see the monograph [89]), and proved to be successful for studying emerging self-organised flow patterns [88]. The theoretical measure framework proves to be also relevant in addressing micro-macro limiting procedures of mean field type [117], where one lets the number of agents going to infinity, while keeping the total mass constant. In this case, one must prove that the *empirical measure*, corresponding to the sum of Dirac measures concentrated at the agents positions, converges to a measure-valued solution of the corresponding macroscopic evolution equation. We recall that a key ingredient in this approach is the use of the *Wasserstein distances* [163], [164]. Indeed, as observed in [147], the usual L^1 spaces are not natural in this context, since they don't guarantee uniqueness of solutions.

This procedure can potentially be extended to more complex configurations, like for example road networks or different classes of interacting agents, or to other application domains, like cell-dynamics.

Another powerful tool we shall consider to deal with micro-macro limits is the so-called **Mean Field Games (MFG)** technique (see the seminal paper [136]). This approach has been recently applied to some of the systems studied by the team, such as traffic flow and cell dynamics. In the context of crowd dynamics, including the case of several populations with different targets, the mean field game approach has been adopted in [72], [73], [98], [135], under the assumption that the individual behavior evolves according to a stochastic process, which gives rise to parabolic equations greatly simplifying the analysis of the system. Besides, a deterministic context is studied in [150], which considers a non-local velocity field. For cell dynamics, in order to take into account the fast processes that occur in the migration-related machinery, a framework such the one developed in [92] to handle games "where agents evolve their strategies according to the best-reply scheme on a much faster time scale than their social configuration variables" may turn out to be suitable. An alternative framework to MFG is also considered. This framework is based on the formulation of -Nash- games

constrained by the **Fokker-Planck** (FP, [62]) partial differential equations that govern the time evolution of the probability density functions -PDF- of stochastic systems and on objectives that may require to follow a given PDF trajectory or to minimize an expectation functional.

3.1.1.3. Non-local flows

Non-local interactions can be described through macroscopic models based on integro-differential equations. Systems of the type

$$\partial_t u + \mathrm{div}_{\mathbf{x}} F(t, \mathbf{x}, u, W) = 0, \qquad t > 0, x \in R^d, d \geq 1, \tag{4}$$

where $u = u(t, \mathbf{x}) \in R^N$, $N \geq 1$ is the vector of conserved quantities and the variable $W = W(t, x, u)$ depends on an integral evaluation of u, arise in a variety of physical applications. Space-integral terms are considered for example in models for granular flows [59], sedimentation [66], supply chains [120], conveyor belts [121], biological applications like structured populations dynamics [146], or more general problems like gradient constrained equations [60]. Also, non-local in time terms arise in conservation laws with memory, starting from [91]. In particular, equations with non-local flux have been recently introduced in traffic flow modeling to account for the reaction of drivers or pedestrians to the surrounding density of other individuals, see [68], [75], [81], [118], [158]. While pedestrians are likely to react to the presence of people all around them, drivers will mainly adapt their velocity to the downstream traffic, assigning a greater importance to closer vehicles. In particular, and in contrast to classical (without integral terms) macroscopic equations, these models are able to display finite acceleration of vehicles through Lipschitz bounds on the mean velocity [68], [118] and lane formation in crossing pedestrian flows.

General analytical results on non-local conservation laws, proving existence and eventually uniqueness of solutions of the Cauchy problem for 1 , can be found in [61] for scalar equations in one space dimension ($N = d = 1$), in [82] for scalar equations in several space dimensions ($N = 1$, $d \geq 1$) and in [55], [83], [87] for multi-dimensional systems of conservation laws. Besides, specific finite volume numerical methods have been developed recently in [55], [118] and [134].

Relying on these encouraging results, we aim to push a step further the analytical and numerical study of non-local models of type 1 , in particular concerning well-posedness of initial - regularity of solutions, boundary value problems and high-order numerical schemes.

3.1.1.4. Uncertainty in parameters and initial-boundary data

Different sources of uncertainty can be identified in PDE models, related to the fact that the problem of interest is not perfectly known. At first, initial and boundary condition values can be uncertain. For instance, in traffic flows, the time-dependent value of inlet and outlet fluxes, as well as the initial distribution of vehicles density, are not perfectly determined [74]. In aerodynamics, inflow conditions like velocity modulus and direction, are subject to fluctuations [124], [145]. For some engineering problems, the geometry of the boundary can also be uncertain, due to structural deformation, mechanical wear or disregard of some details [100]. Another source of uncertainty is related to the value of some parameters in the PDE models. This is typically the case of parameters in turbulence models in fluid mechanics, which have been calibrated according to some reference flows but are not universal [156], [162], or in traffic flow models, which may depend on the type of road, weather conditions, or even the country of interest (due to differences in driving rules and conductors behaviour). This leads to equations with flux functions depending on random parameters [157], [160], for which the mean and the variance of the solutions can be computed using different techniques. Indeed, uncertainty quantification for systems governed by PDEs has become a very active research topic in the last years. Most approaches are embedded in a probabilistic framework and aim at quantifying statistical moments of the PDE solutions, under the assumption that the characteristics of uncertain parameters are known. Note that classical Monte-Carlo approaches exhibit low convergence rate and consequently accurate simulations require huge computational times. In this respect, some enhanced algorithms have been proposed, for example in the balance law framework [143]. Different approaches propose to modify the PDE solvers to account for this probabilistic context, for instance by defining the non-deterministic part of the solution on an orthogonal

basis (Polynomial Chaos decomposition) and using a Galerkin projection [124], [133], [139], [166] or an entropy closure method [95], or by discretizing the probability space and extending the numerical schemes to the stochastic components [54]. Alternatively, some other approaches maintain a fully deterministic PDE resolution, but approximate the solution in the vicinity of the reference parameter values by Taylor series expansions based on first- or second-order sensitivities [151], [162], [165].

Our objective regarding this topic is twofold. In a pure modeling perspective, we aim at including uncertainty quantification in models calibration and validation for predictive use. In this case, the choice of the techniques will depend on the specific problem considered [65]. Besides, we plan to extend previous works on sensitivity analysis [100], [140] to more complex and more demanding problems. In particular, high-order Taylor expansions of the solution (greater than two) will be considered in the framework of the Sensitivity Equation Method [69] (SEM) for unsteady aerodynamic applications, to improve the accuracy of mean and variance estimations. A second targeted topic in this context is the study of the uncertainty related to turbulence closure parameters, in the sequel of [162]. We aim at exploring the capability of the SEM approach to detect a change of flow topology, in case of detached flows. Our ambition is to contribute to the emergence of a new generation of simulation tools, which will provide solution densities rather than values, to tackle real-life uncertain problems. This task will also include a reflection about numerical schemes used to solve PDE systems, in the perspective of constructing a unified numerical framework able to account for exact geometries (isogeometric methods), uncertainty propagation and sensitivity analysis w.r.t. control parameters.

3.1.2. Optimization and control algorithms for systems governed by PDEs

The non-classical models described above are developed in the perspective of design improvement for real-life applications. Therefore, control and optimization algorithms are also developed in conjunction with these models. The focus here is on the methodological development and analysis of optimization algorithms for PDE systems in general, keeping in mind the application domains in the way the problems are mathematically formulated.

3.1.2.1. Sensitivity vs. adjoint equation

Adjoint methods (achieved at continuous or discrete level) are now commonly used in industry for steady PDE problems. Our recent developments [153] have shown that the (discrete) adjoint method can be efficiently applied to cost gradient computations for time-evolving traffic flow on networks, thanks to the special structure of the associated linear systems and the underlying one dimensionality of the problem. However, this strategy is questionable for more complex (e.g. 2D/3D) unsteady problems, because it requires sophisticated and time-consuming check-pointing and/or re-computing strategies [63], [119] for the backward time integration of the adjoint variables. The sensitivity equation method (SEM) offers a promising alternative [99], [129], if the number of design parameters is moderate. Moreover, this approach can be employed for other goals, like fast evaluation of neighboring solutions or uncertainty propagation [100].
Regarding this topic, we intend to apply the continuous sensitivity equation method to challenging problems. In particular, in aerodynamics, multi-scale turbulence models like Large-Eddy Simulation (LES) [155] , Detached-Eddy Simulation (DES) [159] or Organized-Eddy Simulation (OES) [70], are more and more employed to analyse the unsteady dynamics of the flows around bluff-bodies, because they have the ability to compute the interactions of vortices at different scales, contrary to classical Reynolds-Averaged Navier-Stokes models. However, their use in design optimization is tedious, due to the long time integration required. In collaboration with turbulence specialists (M. Braza, CNRS - IMFT), we aim at developing numerical methods for effective sensitivity analysis in this context, and apply them to realistic problems, like the optimization of active flow control devices. Note that the use of SEM allows computing cost functional gradients at any time, which permits to construct new gradient-based optimization strategies like instantaneous-feedback method [131] or multiobjective optimization algorithm (see section below).

3.1.2.2. Multi-objective descent algorithms for multi-disciplinary, multi-point, unsteady optimization or robust-design

In differentiable optimization, multi-disciplinary, multi-point, unsteady optimization or robust-design can all be formulated as multi-objective optimization problems. In this area, we have proposed the *Multiple-Gradient*

Descent Algorithm (MGDA) to handle all criteria concurrently [102] [103]. Originally, we have stated a principle according which, given a family of local gradients, a descent direction common to all considered objective-functions simultaneously is identified, assuming the Pareto-stationarity condition is not satisfied. When the family is linearly-independent, we dispose of a direct algorithm. Inversely, when the family is linearly-dependent, a quadratic-programming problem should be solved. Hence, the technical difficulty is mostly conditioned by the number m of objective functions relative to the search space dimension n. In this respect, the basic algorithm has recently been revised [104] to handle the case where $m > n$, and even $m \gg n$, and is currently being tested on a test-case of robust design subject to a periodic time-dependent Navier-Stokes flow.

The multi-point situation is very similar and, being of great importance for engineering applications, will be treated at large.

Moreover, we intend to develop and test a new methodology for robust design that will include uncertainty effects. More precisely, we propose to employ MGDA to achieve an effective improvement of all criteria simultaneously, which can be of statistical nature or discrete functional values evaluated in confidence intervals of parameters. Some recent results obtained at ONERA [148] by a stochastic variant of our methodology confirm the viability of the approach. A PhD thesis has also been launched at ONERA/DADS.

Lastly, we note that in situations where gradients are difficult to evaluate, the method can be assisted by a meta-model [168].

3.1.2.3. Bayesian Optimization algorithms for efficient computation of general equilibria

Bayesian Optimization (BO) relies on Gaussian processes, which are used as emulators (or surrogates) of the black-box model outputs based on a small set of model evaluations. Posterior distributions provided by the Gaussian process are used to design acquisition functions that guide sequential search strategies that balance between exploration and exploitation. Such approaches have been transposed to frameworks other than optimization, such as uncertainty quantification. Our aim is to investigate how the BO apparatus can be applied to the search of general game equilibria, and in particular the classical Nash equilibrium (NE). To this end, we propose two complementary acquisition functions, one based on a greedy search approach and one based on the Stepwise Uncertainty Reduction paradigm [112]. Our proposal is designed to tackle derivative-free, expensive models, hence requiring very few model evaluations to converge to the solution.

3.1.2.4. Decentralized strategies for inverse problems

Most if not all the mathematical formulations of inverse problems (a.k.a. reconstruction, identification, data recovery, non destructive engineering,...) are known to be ill posed in the Hadamard sense. Indeed, in general, inverse problems try to fulfill (minimize) two or more very antagonistic criteria. One classical example is the Tikhonov regularization, trying to find artificially smoothed solutions close to naturally non-smooth data.

We consider here the theoretical general framework of parameter identification coupled to (missing) data recovery. Our aim is to design, study and implement algorithms derived within a game theoretic framework, which are able to find, with computational efficiency, equilibria between the "identification related players" and the "data recovery players". These two parts are known to pose many challenges, from a theoretical point of view, like the identifiability issue, and from a numerical one, like convergence, stability and robustness problems. These questions are tricky [56] and still completely open for systems like e.g. coupled heat and thermoelastic joint data and material detection.

BONUS Project-Team

3. Research Program

3.1. Decomposition-based Optimization

Given the large scale of the targeted optimization problems in terms of the number of variables and objectives, their decomposition into simplified and loosely coupled or independent subproblems is essential to raise the challenge of scalability. The first line of research is to *investigate the decomposition approach in the two spaces and their combination, as well as their implementation on ultra-scale architectures*. The motivation of the decomposition is twofold: first, the decomposition allows the parallel resolution of the resulting subproblems on ultra-scale architectures. Here also several issues will be addressed: the definition of the subproblems, their coding to allow their efficient communication and storage (checkpointing), their assignment to processing cores etc. Second, decomposition is necessary for solving large problems that cannot be solved (efficiently) using traditional algorithms. Indeed, for instance with the popular NSGA-II algorithm the number of non-dominated solutions [0] increases drastically with the number of objectives leading to a very slow convergence to the Pareto Front [0]. Therefore, decomposition-based techniques are gaining a growing interest. The objective of BONUS is to *investigate various decomposition schema and cooperation protocols between the subproblems* resulting from the decomposition to generate efficiently global solutions of good quality. Several challenges have to be addressed: (1) how to define the subproblems (decomposition strategy), (2) how to solve them to generate local solutions (local rules), and (3) how to combine these latter with those generated by other subproblems and how to generate global solutions (cooperation mechanism), and (4) how to combine decomposition strategies in more than one space (hybridization strategy)? These challenges, which are in the line with the CIS Task Force [0] on decomposition will be addressed in the decision as well as in the objective space.

The *decomposition in the decision space* can be performed following different ways according to the problem at hand. Two major categories of decomposition techniques can be distinguished: the first one consists in *breaking down the high-dimensional decision vector* into lower-dimensional and easier-to-optimize blocks of variables. The major issue is how to define the subproblems (blocks of variables) and their cooperation protocol: randomly *vs.* using some learning (e.g. separability analysis), statically *vs.* adaptively etc. *The decomposition in the decision space can also be guided by the type of variables i.e. discrete vs. continuous*. The discrete and continuous parts are optimized separately using cooperative hybrid algorithms [48]. *The major issue of this kind of decomposition is the presence of categorial variables in the discrete part [44]. The* BONUS *team is addressing this issue, rarely investigated in the literature*, within the context of vehicle aerospace engineering design. The second category consists in the *decomposition according to the ranges of the decision variables*. For continuous problems, the idea consists in iteratively subdividing the search (e.g. design) space into subspaces (hyper-rectangles, intervals etc.) and select those that are most likely to produce the lowest objective function value. *Existing approaches meet increasing difficulty with an increasing number of variables and are often applied to low-dimensional problems. We are investigating this scalability challenge* (e.g. [10]). *For discrete problems, the major challenge is to find a coding (mapping) of the search space to a decomposable entity*. We have proposed an interval-based coding of the permutation space for solving big permutation problems. The approach opens perspectives we are investigating [7], in terms of ultra-scale parallelization, application to multi-permutation problems and hybridization with metaheuristics.

[0] A solution x dominates another solution y if x is better than y for all objectives and there exists at least one objective for which x is strictly better than y.
[0] The Pareto Front is the set of non-dominated solutions.
[0] IEEE CIS Task Force, created in 2017 on Decomposition-based Techniques in Evolutionary Computation.

The *decomposition in the objective space* consists in breaking down an original Many-objective problem (MaOP) into a set of cooperative single-objective subproblems (SOPs). The decomposition strategy requires the careful definition of a scalarizing (aggregation) function and its weighting vectors (each of them corresponds to a separate SOP) to guide the search process towards the best regions. Several scalarizing functions have been proposed in the literature including weighted sum, weighted Tchebycheff, vector angle distance scaling etc. These functions are widely used but they have their limitations. For instance, using weighted Tchebycheff might do harm diversity maintenance and weighted sum is inefficient when it comes to deal with nonconvex Pareto Fronts [40]. Defining a scalarizing function well-suited to the MaOP at hand is therefore a difficult and still an open question being investigated in BONUS [6], [5]. Studying/defining various functions and in-depth analyzing them to better understand the differences between them is required. Regarding the weighting vectors that determine the search direction, their efficient setting is also a key and open issue. They dramatically affect in particular the diversity performance. Their setting rises two main issues: how to determine their number according to the available computational ressources? when (statically or adaptively) and how to determine their values? *Weight adaptation is one of our main concerns that we are addressing especially from a distributed perspective.* They correspond to the main scientific objectives targeted by our bilateral ANR-RGC BigMO project with City University (Hong Kong). The other challenges pointed out in the beginning of this section concern the way to solve locally the SOPs resulting from the decomposition of a MaOP and the mechanism used for their cooperation to generate global solutions. To deal with these challenges, our approach is to design the decomposition strategy and cooperation mechanism keeping in mind the parallel and/or distributed solving of the SOPs. Indeed, we favor the local neighborhood-based mating selection and replacement to minimize the network communication cost while allowing an effective resolution [5]. The major issues here are how to define the neighborhood of a subproblem and how to cooperatively update the best-known solution of each subproblem and its neighbors.

To sum up, the objective of the BONUS *team is to come up with scalable decomposition-based approaches in the decision and objective spaces. In the decision space, a particular focus will be put on high dimensionality and mixed-continuous variables which have received little interest in the literature. We will particularly continue to investigate at larger scales using ultra-scale computing the interval-based (discrete) and fractal-based (continuous) approaches. We will also deal with the rarely addressed challenge of mixed-continuous including categorial variables (collaboration with ONERA). In the objective space, we will investigate parallel ultra-scale decomposition-based many-objective optimization with ML-based adaptive building of scalarizing functions. A particular focus will be put on the state-of-the-art MOEA/D algorithm. This challenge is rarely addressed in the literature which motivated the collaboration with the designer of MOEA/D (bilateral ANR-RGC BigMO project with City University, Hong Kong). Finally, the joint decision-objective decomposition, which is still in its infancy [50], is another challenge of major interest.*

3.2. Machine Learning-assisted Optimization

The Machine Learning (ML) approach based on metamodels (or surrogates) is commonly used, and also adopted in BONUS, to assist optimization in tackling BOPs characterized by time-demanding objective functions. The second line of research of BONUS is focused on ML-aided optimization to raise the challenge of expensive functions of BOPs using surrogates but also to assist the two other research lines (decomposition-based and ultra-scale optimization) in dealing with the other challenges (high dimensionality and scalability). Several issues have been identified to make efficient and effective surrogate-assisted optimization. First, infill criteria have to be carefully defined to adaptively select the adequate sample points (in terms of surrogate precision and solution quality). The challenge is to find the best trade-off between exploration and exploitation to efficiently refine the surrogate and guide the optimization process toward the best solutions. The most popular infill criterion is probably the *Expected Improvement* (EI) [43] which is based on the expected values of sample points but also and importantly on their variance. This latter is inherently determined in the kriging model, this is why it is used in the state-of-the-art *efficient global optimization* (EGO) algorithm [43]. However, such crucial information is not provided in all surrogate models (e.g. Artificial Neural Networks) and needs to be derived. In BONUS, we are currently investigating this issue. Second, it is known that surrogates allow one to reduce the computational burden for solving BOPs with time-demanding function(s). However,

using parallel computing as a complementary way is often recommended and cited as a perspective in the conclusions of related publications. Nevertheless, *despite being of critical importance parallel surrogate-assisted optimization is weakly addressed in the literature*. For instance, in the introduction of the survey proposed in [42] it is warned that because the area is not mature yet the paper is more focused on the potential of the surveyed approaches than on their relative efficiency. *Parallel computing is required at different levels that we are investigating*.

Another issue with surrogate-assisted optimization is related to high dimensionality in decision as well as in objective space: it is often applied to low-dimensional problems. *The joint use of decomposition, surrogates and massive parallelism is an efficient approach to deal with high dimensionality. This approach adopted in* BONUS *has received little effort in the literature.* In BONUS, we are considering a generic framework in order to enable a flexible coupling of existing surrogate models within the state-of-the-art decomposition-based algorithm MOEA/D. This is a first step in leveraging the applicability of efficient global optimization into the multi-objective setting through parallel decomposition. Another issue which is a consequence of high dimensionality is the mixed (discrete-continuous) nature of decision variables which is frequent in real-world applications (e.g. engineering design). *While surrogate-assisted optimization is widely applied in the continuous setting it is rarely addressed in the literature in the discrete-continuous framework.* In [44], we have identified different ways to deal with this issue that we are investigating. Non-stationary functions frequent in real-world applications (see Section 4.1) is another major issue we are addressing using the concept of deep GP.

Finally, as quoted in the beginning of this section, ML-assisted optimization is mainly used to deal with BOPs with expensive functions but it will also be investigated for other optimization tasks. Indeed, ML will be useful to assist the decomposition process. In the decision space, it will help to perform the separability analysis (understanding of the interactions between variables) to decompose the vector of variables. In the objective space, ML will be useful to assist a decomposition-based many-objective algorithm in dynamically selecting a scalarizing function or updating the weighting vectors according to their performances in the previous steps of the optimization process [5]. Such a data-driven ML methodology would allow us to understand what makes a problem difficult or an optimization approach efficient, to predict the algorithm performance [4], to select the most appropriate algorithm configuration [8], and to adapt and improve the algorithm design for unknown optimization domains and instances. Such an autonomous optimization approach would adaptively adjust its internal mechanisms in order to tackle cross-domain BOPs.

In a nutshell, to deal with expensive optimization the BONUS *team will investigate the surrogate-based ML approach with the objective to efficiently integrate surrogates in the optimization process. The focus will especially be put on high dimensionality (e.g. using decomposition) with mixed discrete-continuous variables which is rarely investigated. The kriging metamodel (Gaussian Process or GP) will be considered in particular for engineering design (for more reliability) addressing the above issues and other major ones including mainly non stationarity (using emerging deep GP) and ultra-scale parallelization (highly needed by the community). Indeed, a lot of work has been reported on deep neural networks (deep learning) surrogates but not on the others including (deep) GP. On the other hand, ML will be used to assist decomposition: importance/interaction between variables in the decision space, dynamic building (selection of scalarizing functions, weight update etc.) of scalarizing functions in the objective space etc.*

3.3. Ultra-scale Optimization

The third line of our research program that accentuates our difference from other (project-)teams of the related Inria scientific theme is the ultra-scale optimization. *This research line is complementary to the two others, which are sources of massive parallelism* and with which it should be combined to solve BOPs. Indeed, ultra-scale computing is necessary for the effective resolution of the large amount of subproblems generated by decomposition of BOPs, parallel evaluation of simulation-based fitness and metamodels etc. These sources of parallelism are attractive for solving BOPs and are natural candidates for ultra-scale supercomputers [0].

[0] In the context of BONUS, supercomputers are composed of several massively parallel processing nodes (inter-node parallelism) including multi-core processors and GPUs (intra-node parallelism).

However, their efficient use raises a big challenge consisting in managing efficiently a massive amount of irregular tasks on supercomputers with multiple levels of parallelism and heterogeneous computing resources (GPU, multi-core CPU with various architectures) and networks. Raising such challenge requires to tackle three major issues, scalability, heterogeneity and fault-tolerance, discussed in the following.

The *scalability* issue requires, on the one hand, the definition of scalable data structures for efficient storage and management of the tremendous amount of subproblems generated by decomposition [46]. On the other hand, achieving extreme scalability requires also the optimization of communications (in number of messages, their size and scope) especially at the inter-node level. For that, we target the design of asynchronous locality-aware algorithms as we did in [41], [49]. In addition, efficient mechanisms are needed for granularity management and coding of the work units stored and communicated during the resolution process.

Heterogeneity means harnessing various resources including multi-core processors within different architectures and GPU devices. The challenge is therefore to design and implement hybrid optimization algorithms taking into account the difference in computational power between the various resources as well as the resource-specific issues. On the one hand, to deal with the heterogeneity in terms of computational power, we adopt in BONUS the dynamic load balancing approach based on the Work Stealing (WS) asynchronous paradigm [0] at the inter-node as well as at the intra-node level. We have already investigated such approach, with various victim selection and work sharing strategies in [49], [7]. On the other hand, hardware resource specific-level optimization mechanisms are required to deal with related issues such as thread divergence and memory optimization on GPU, data sharing and synchronization, cache locality, and vectorization on multi-core processors etc. These issues have been considered separately in the literature including our works [9], [1]. Indeed, in most of existing works related to GPU-accelerated optimization only a single CPU core is used. This leads to a huge resource wasting especially with the increase of the number of processing cores integrated into modern processors. Using jointly the two components raises additional issues including data and work partitioning, the optimization of CPU-GPU data transfers etc.

Another issue the scalability induces is the *increasing probability of failures* in modern supercomputers [47]. Indeed, with the increase of their size to millions of processing cores their Mean-Time Between Failures (MTBF) tends to be shorter and shorter [45]. Failures may have different sources including hardware and software faults, silent errors etc. In our context, we consider failures leading to the loss of work unit(s) being processed by some thread(s) during the resolution process. The major issue, which is particularly critical in exact optimization, is how to recover the failed work units to ensure a reliable execution. Such issue is tackled in the literature using different approaches: algorithm-based fault tolerance, checkpoint/restart (CR), message logging and redundancy. The CR approach can be system-level, library/user-level or application-level. Thanks to its efficiency in terms of memory footprint, adopted in BONUS [2], the application-level approach is commonly and widely used in the literature. This approach raises several issues mainly: (1) which critical information defines the state of the work units and allows to resume properly their execution? (2) when, where and how (using which data structures) to store it efficiently? (3) how to deal with the two other issues: scalability and heterogeneity?
The last but not least major issue which is another roadblock to exascale is the programming of massive-scale applications for modern supercomputers. *On the path to exascale, we will investigate the programming environments and execution supports able to deal with exascale challenges: large numbers of threads, heterogeneous resources etc.* Various exascale programming approaches are being investigated by the parallel computing community and HPC builders: extending existing programming languages (e.g. DSL-C++) and environments/libraries (MPI+X etc.), proposing new solutions including mainly Partitioned Global Address Space (PGAS)-based environments (Chapel, UPC, X10 etc.). It is worth noting here that our objective is not to develop a programming environment nor a runtime support for exascale computing. Instead, we aim to collaborate with the research teams (inside or outside Inria) having such objective.

[0] A WS mechanism is mainly defined by two components: a victim selection strategy which selects the processing core to be stolen and a work sharing policy which determines the part and amount of the work unit to be given to the thief upon WS request.

To sum up, we put the focus on the design and implementation of efficient big optimization algorithms dealing jointly (uncommon in parallel optimization) with the major issues of ultra-scale computing mainly the scalability up to millions of cores using scalable data structures and asynchronous locality-aware work stealing, heterogeneity addressing the multi-core and GPU-specific issues and those related to their combination, and scalable GPU-aware fault tolerance. A strong effort will be devoted to this latter challenge, for the first time to the best of our knowledge, using application-level checkpoint/restart approach to deal with failures.

CAGE Project-Team

3. Research Program

3.1. Research domain

The activities of CAGE are part of the research in the wide area of control theory. This nowadays mature discipline is still the subject of intensive research because of its crucial role in a vast array of applications.

More specifically, our contributions are in the area of **mathematical control theory**, which is to say that we are interested in the analytical and geometrical aspects of control applications. In this approach, a control system is modeled by a system of equations (of many possible types: ordinary differential equations, partial differential equations, stochastic differential equations, difference equations,...), possibly not explicitly known in all its components, which are studied in order to establish qualitative and quantitative properties concerning the actuation of the system through the control.

Motion planning is, in this respect, a cornerstone property: it denotes the design and validation of algorithms for identifying a control law steering the system from a given initial state to (or close to) a target one. Initial and target positions can be replaced by sets of admissible initial and final states as, for instance, in the motion planning task towards a desired periodic solution. Many specifications can be added to the pure motion planning task, such as robustness to external or endogenous disturbances, obstacle avoidance or penalization criteria. A more abstract notion is that of **controllability**, which denotes the property of a system for which any two states can be connected by a trajectory corresponding to an admissible control law. In mathematical terms, this translates into the surjectivity of the so-called **end-point map**, which associates with a control and an initial state the final point of the corresponding trajectory. The analytical and topological properties of endpoint maps are therefore crucial in analyzing the properties of control systems.

One of the most important additional objective which can be associated with a motion planning task is **optimal control**, which corresponds to the minimization of a cost (or, equivalently, the maximization of a gain) [156]. Optimal control theory is clearly deeply interconnected with calculus of variations, even if the non-interchangeable nature of the time-variable results in some important specific features, such as the occurrence of **abnormal extremals** [120]. Research in optimal control encompasses different aspects, from numerical methods to dynamic programming and non-smooth analysis, from regularity of minimizers to high order optimality conditions and curvature-like invariants.

Another domain of control theory with countless applications is **stabilization**. The goal in this case is to make the system converge towards an equilibrium or some more general safety region. The main difference with respect to motion planning is that here the control law is constructed in feedback form. One of the most important properties in this context is that of **robustness**, i.e., the performance of the stabilization protocol in presence of disturbances or modeling uncertainties. A powerful framework which has been developed to take into account uncertainties and exogenous non-autonomous disturbances is that of hybrid and switched systems [159], [119], [147]. The central tool in the stability analysis of control systems is that of **control Lyapunov function**. Other relevant techniques are based on algebraic criteria or dynamical systems. One of the most important stability property which is studied in the context of control system is **input-to-state stability** [143], which measures how sensitive the system is to an external excitation.

One of the areas where control applications have nowadays the most impressive developments is in the field of **biomedicine and neurosciences**. Improvements both in modeling and in the capability of finely actuating biological systems have concurred in increasing the popularity of these subjects. Notable advances concern, in particular, identification and control for biochemical networks [137] and models for neural activity [106]. Therapy analysis from the point of view of optimal control has also attracted a great attention [140].

Biological models are not the only one in which stochastic processes play an important role. Stock-markets and energy grids are two major examples where optimal control techniques are applied in the non-deterministic setting. Sophisticated mathematical tools have been developed since several decades to allow for such extensions. Many theoretical advances have also been required for dealing with complex systems whose description is based on **distributed parameters** representation and **partial differential equations**. Functional analysis, in particular, is a crucial tool to tackle the control of such systems [153].

Let us conclude this section by mentioning another challenging application domain for control theory: the decision by the European Union to fund a flagship devoted to the development of quantum technologies is a symptom of the role that quantum applications are going to play in tomorrow's society. **Quantum control** is one of the bricks of quantum engineering, and presents many peculiarities with respect to standard control theory, as a consequence of the specific properties of the systems described by the laws of quantum physics. Particularly important for technological applications is the capability of inducing and reproducing coherent state superpositions and entanglement in a fast, reliable, and efficient way [107].

3.2. Scientific foundations

At the core of the scientific activity of the team is the **geometric control** approach, that is, a distinctive viewpoint issued in particular from (elementary) differential geometry, to tackle questions of controllability, observability, optimal control... [70], [111]. The emphasis of such a geometric approach to control theory is put on intrinsic properties of the systems and it is particularly well adapted to study nonlinear and nonholonomic phenomena.

One of the features of the geometric control approach is its capability of exploiting **symmetries and intrinsic structures** of control systems. Symmetries and intrinsic structures can be used to characterize minimizing trajectories, prove regularity properties and describe invariants. An egregious example is given by mechanical systems, which inherently exhibit Lagrangian/Hamiltonian structures which are naturally expressed using the language of symplectic geometry [93]. The geometric theory of quantum control, in particular, exploits the rich geometric structure encoded in the Schrödinger equation to engineer adapted control schemes and to characterize their qualitative properties. The Lie–Galerkin technique that we proposed starting from 2009 [96] builds on this premises in order to provide powerful tests for the controllability of quantum systems defined on infinite-dimensional Hilbert spaces.

Although the focus of geometric control theory is on qualitative properties, its impact can also be disruptive when it is used in combination with quantitative analytical tools, in which case it can dramatically improve the computational efficiency. This is the case in particular in optimal control. Classical optimal control techniques (in particular, Pontryagin Maximum Principle, conjugate point theory, associated numerical methods) can be significantly improved by combining them with powerful modern techniques of geometric optimal control, of the theory of numerical continuation, or of dynamical system theory [152], [139]. Geometric optimal control allows the development of general techniques, applying to wide classes of nonlinear optimal control problems, that can be used to characterize the behavior of optimal trajectories and in particular to establish regularity properties for them and for the cost function. Hence, geometric optimal control can be used to obtain powerful optimal syntheses results and to provide deep geometric insights into many applied problems. Numerical optimal control methods with geometric insight are in particular important to handle subtle situations such as rigid optimal paths and, more generally, optimal syntheses exhibiting abnormal minimizers.

Optimal control is not the only area where the geometric approach has a great impact. Let us mention, for instance, motion planning, where different geometric approaches have been developed: those based on the **Lie algebra** associated with the control system [132], [122], those based on the differentiation of nonlinear flows such as the **return method** [101], [100], and those exploiting the **differential flatness** of the system [105].

Geometric control theory is not only a powerful framework to investigate control systems, but also a useful tool to model and study phenomena that are not *a priori* control-related. Two occurrences of this property play an important role in the activities of CAGE:

- geometric control theory as a tool to investigate properties of mathematical structures;

- geometric control theory as a modeling tool for neurophysical phenomena and for synthesizing biomimetic algorithms based on such models.

Examples of the first type, concern, for instance, hypoelliptic heat kernels [68] or shape optimization [76]. Examples of the second type are inactivation principles in human motricity [79] or neurogeometrical models for image representation of the primary visual cortex in mammals [90].

A particularly relevant class of control systems, both from the point of view of theory and applications, is characterized by the linearity of the controlled vector field with respect to the control parameters. When the controls are unconstrained in norm, this means that the admissible velocities form a distribution in the tangent bundle to the state manifold. If the distribution is equipped with a point-dependent quadratic form (encoding the cost of the control), the resulting geometrical structure is said to be **sub-Riemannian**. Sub-Riemannian geometry appears as the underlying geometry of nonlinear control systems: in a similar way as the linearization of a control system provides local informations which are readable using the Euclidean metric scale, sub-Riemannian geometry provides an adapted non-isotropic class of lenses which are often much more informative. As such, its study is fundamental for control design. The importance of sub-Riemannian geometry goes beyond control theory and it is an active field of research both in differential geometry [130], geometric measure theory [72] and hypoelliptic operator theory [82].

The geometric control approach has historically been related to the development of finite-dimensional control theory. However, its impact in the analysis of distributed parameter control systems and in particular systems of controlled partial differential equations has been growing in the last decades, complementing analytical and numerical approaches, providing dynamical, qualitative and intrinsic insight [99]. CAGE's ambition is to be at the core of this development in the years to come.

CAGIRE Project-Team

3. Research Program

3.1. The scientific context

3.1.1. Computational fluid mechanics: modeling or not before discretizing ?

A typical continuous solution of the Navier-Stokes equations at sufficiently large values of the Reynolds number is governed by a wide spectrum of temporal and spatial scales closely connected with the turbulent nature of the flow. The term deterministic chaos employed by Frisch in his enlightening book [46] is certainly conveying most adequately the difficulty in analyzing and simulating this kind of flows. The broadness of the turbulence spectrum is directly controlled by the Reynolds number defined as the ratio between the inertial forces and the viscous forces. This number is not only useful to determine the transition from a laminar to a turbulent flow regime, it also indicates the range of scales of fluctuations that are present in the flow under consideration. Typically, for the velocity field and far from solid walls, the ratio between the largest scale (the integral length scale) and the smallest one (Kolmogorov scale) is proportional to $Re_t^{3/4}$ per dimension, where $Re_t^{3/4}$ is the turbulent Reynolds number, based on the length and velocity scales of the largest turbulent eddies. In addition, for internal flows, viscous effects near the solid walls yield a scaling proportional to Re_τ per dimension, where Re_τ is the friction Reynolds number. The smallest scales play a crucial role in the dynamics of the largest ones, which implies that an accurate framework for the computation of turbulent flows must take into account all the scales, which can lead to unrealistic computational costs in real-world applications. Thus, the usual practice to deal with turbulent flows is to choose between an a priori modeling (in most situations) or not (low Re number and rather simple configurations) before proceeding to the discretization step, followed by the simulation itself. If a modeling phase is on the agenda, then one has to choose again among the above-mentioned variety of approaches. The different simulation options and their date of availability for high-Reynolds-number applications are illustrated in Fig. 1 : simulation of turbulent flows can be achieved either by directly solving the Navier-Stokes equations (DNS) or by first applying to the equations a statistical averaging (RANS), a spatial filtering (LES), or a combination of these two operators (hybrid RANS/LES). The new terms brought about by the operator have to be modeled. From a computational point of view, the RANS approach is the least demanding, which explains why historically it has been the workhorse in both the academic and the industrial sectors, and it remains the standard approach nowadays for industrial design, except for very specific applications. It has permitted quite a substantial progress in the understanding of various phenomena such as turbulent combustion or heat transfer. Its inherent inability to provide a time-dependent information has led to promote in the last decade the recourse to either LES or DNS to supplement if not replace RANS. By simulating the large scale structures while modeling the smallest ones, assumed more isotropic, LES proved to be quite a breakthrough to fully take advantage of the increasing power of computers to study complex flow configurations. At the same time, DNS was gradually applied to geometries of increasing complexity (channel flows with values of Re_τ multiplied by 45 during the last 30 years, jets, turbulent premixed flames, among many others), and proved to be a formidable tool to **(i)** improve our knowledge on turbulent flows and **(ii)** test (i.e., validate or invalidate) and improve the modeling hypotheses inherently associated to the RANS and LES approaches. From a numerical point of view, due to the steady nature of the RANS equations, numerical accuracy is generally not ensured via the use of high-order schemes, but rather on careful grid convergence studies. In contrast, the high computational cost of LES or DNS makes necessary the use of highly-accurate numerical schemes in order to optimize the use of computational resources.

To the noticeable exception of the hybrid RANS-LES modeling, which is not yet accepted as a reliable tool for industrial design, as mentioned in the preamble of the Go4hybrid European program [0], a turbulence model represents turbulent mechanisms in the same way in the whole flow. Thus, depending on its intrinsic strengths

[0] https://cordis.europa.eu/project/rcn/109107/factsheet/en

and weaknesses, accuracy will be a rather volatile quantity, strongly dependent on the flow configuration. For instance, RANS is perfectly suited to attached boundary layers, but exhibits severe limitations in massively-separated flow regions. Therefore, the turbulence modeling and industrial design communities waver between the desire to continue to rely on the RANS approach, which is unrivaled in terms of computational cost, but is still not able to accurately represent all the complex phenomena; and the temptation to switch to LES, which outperforms RANS in many situations, but is prohibitively expensive in high-Reynolds number wall-bounded flows. In order to account for the limitations of the two approaches and to combine them for significantly improving the overall performance of the models, the hybrid RANS-LES approach has emerged during the last two decades as a viable, intermediate way, and we are definitely inscribing our project in this innovative field of research, with an original approach though, based on temporal filtering (Hybrid temporal LES, HTLES) rather than spatial filtering, and a systematic and progressive validation process against experimental data produced by the team.

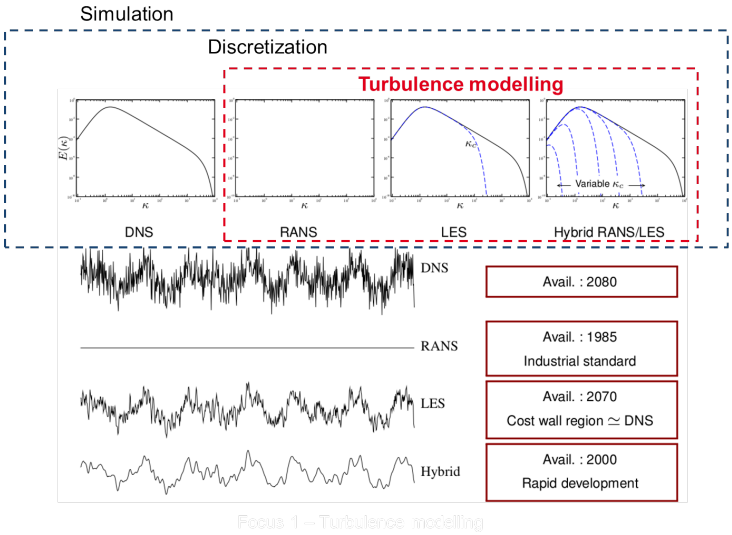


Figure 1. Schematic view of the different nested steps for turbulent flow simulation: from DNS to hybrid RANS-LES. The approximate dates at which the different approaches are or will be routinely used in the industry are indicated in the boxes on the right (extrapolations based on the present rate of increase in computer performances).

3.1.2. Computational fluid mechanics: high order discretization on unstructured meshes and efficient methods of solution

All the methods considered in the project are mesh-based methods: the computational domain is divided into cells, that have an elementary shape: triangles and quadrangles in two dimensions, and tetrahedra, hexahedra, pyramids, and prisms in three dimensions. If the cells are only regular hexahedra, the mesh is said to be structured. Otherwise, it is said to be unstructured. If the mesh is composed of more than one sort of elementary shape, the mesh is said to be hybrid. In the project, the numerical strategy is based on discontinuous Galerkin methods. These methods were introduced by Reed and Hill [57] and first studied by Lesaint and Raviart [53]. The extension to the Euler system with explicit time integration was mainly led by Shu, Cockburn and their collaborators. The steps of time integration and slope limiting were similar to high-order ENO schemes, whereas specific constraints given by the finite-element nature of the scheme were gradually solved for scalar conservation laws [41], [40], one dimensional systems [39], multidimensional scalar conservation laws [38], and multidimensional systems [42]. For the same system, we can also cite the work of [45], [50], which is

slightly different: the stabilization is made by adding a nonlinear term, and the time integration is implicit. In contrast to continuous Galerkin methods, the discretization of diffusive operators is not straightforward. This is due to the discontinuous approximation space, which does not fit well with the space function in which the diffusive system is well posed. A first stabilization was proposed by Arnold [31]. The first application of discontinuous Galerkin methods to Navier-Stokes equations was proposed in [36] by mean of a mixed formulation. Actually, this first attempt led to a non-compact computational stencil, and was later proved to be unstable. A compactness improvement was made in [37], which was later analyzed, and proved to be stable in a more unified framework [32]. The combination with the $k-\omega$ RANS model was made in [35]. As far as Navier-Stokes equations are concerned, we can also cite the work of [48], in which the stabilization is closer to the one of [32], the work of [54] on local time stepping, or the first use of discontinuous Galerkin methods for direct numerical simulation of a turbulent channel flow done in [43]. Discontinuous Galerkin methods became very popular because:

- They can be developed for any order of approximation.
- The computational stencil of one given cell is limited to the cells with which it has a common face. This stencil does not depend on the order of approximation. This is a pro, compared for example with high-order finite volumes, for which the number of neighbors required increases with the order of approximation.
- They can be developed for any kind of mesh, structured, unstructured, but also for aggregated grids [34]. This is a pro compared not only with finite-difference schemes, which can be developed only on structured meshes, but also compared with continuous finite-element methods, for which the definition of the approximation basis is not clear on aggregated elements.
- p-adaptivity is easier than with continuous finite elements, because neighboring elements having a different order are only weakly coupled.
- Upwinding is as natural as for finite volumes methods, which is a benefit for hyperbolic problems.
- As the formulation is weak, boundary conditions are naturally weakly formulated. This is a benefit compared with strong formulations, for example point centered formulation when a point is at the intersection of two kinds of boundary conditions.

For concluding this section, there already exists numerical schemes based on the discontinuous Galerkin method, which proved to be efficient for computing compressible viscous flows. Nevertheless, there remain many things to be improved, which include: efficient shock capturing methods for supersonic flows, high-order discretization of curved boundaries, low-Mach-number behavior of these schemes and combination with second-moment RANS closures. Another aspect that deserves attention is the computational cost of discontinuous Galerkin methods, due to the accurate representation of the solution, calling for a particular care of implementation for being efficient. We believe that this cost can be balanced by the strong memory locality of the method, which is an asset for porting on emerging many-core architectures.

3.1.3. Experimental fluid mechanics: a relevant tool for physical modeling and simulation development

With the considerable and constant development of computer performance, many people were thinking at the turn of the 21st century that in the short term, CFD would replace experiments, considered as too costly and not flexible enough. Simply flipping through scientific journals such as Journal of Fluid Mechanics, Combustion and Flame, Physics of Fluids or Journal of Computational Physics or through websites such that of Ercoftac [0] is sufficient to convince oneself that the recourse to experiments to provide either a quantitative description of complex phenomena or reference values for the assessment of the predictive capabilities of models and simulations is still necessary. The major change that can be noted though concerns the content of the interaction between experiments and CFD (understood in the broad sense). Indeed, LES or DNS assessment calls for the experimental determination of temporal and spatial turbulent scales, as well as time-resolved measurements and determination of single or multi-point statistical properties of the velocity field. Thus, the team methodology incorporates from the very beginning an experimental component that is operated in strong interaction with the modeling and simulation activities.

[0] http://www.ercoftac.org

3.2. Research directions

3.2.1. Boundary conditions

3.2.1.1. Generating synthetic turbulence

A crucial point for any multi-scale simulation able to locally switch (in space or time) from a coarse to a fine level of description of turbulence, is the enrichment of the solution by fluctuations as physically meaningful as possible. Basically, this issue is an extension of the problem of the generation of realistic inlet boundary conditions in DNS or LES of subsonic turbulent flows. In that respect, the method of anisotropic linear forcing (ALF) we have developed in collaboration with EDF proved very encouraging, by its efficiency, its generality and simplicity of implementation. So, it seems natural, on the one hand, to extend this approach to the compressible framework and to implement it in AeroSol. On the other hand, we shall concentrate (in cooperation with EDF R&D in Chatou in the framework of a the CIFRE PhD of V. Duffal) on the theoretical link between the local variations of the scale of description of turbulence (e.g. a sudden variations in the size of the time filter) and the intensity of the ALF forcing, transiently applied to promote the development of missing fluctuating scales.

3.2.1.2. Stable and non reflecting boundary conditions

In aerodynamics, and especially for subsonic computations, handling inlet and outlet boundary conditions is a difficult issue. A significant amount of work has already been performed for second-order schemes for Navier-Stokes equations, see [56], [59] and the huge number of papers citing it. On the one hand, we believe that decisive improvements are necessary for higher-order schemes: indeed, the less dissipative the scheme is, the worse impact have the spurious reflections. For this purpose, we will first concentrate on the linearized Navier-Stokes system, and analyze the way to impose boundary conditions in a discontinuous Galerkin framework with a similar approach as in [47]. We will also try to extend the work of [60], which deals with Euler equations, to the Navier-Stokes equations.

3.2.2. Turbulence models and model agility

3.2.2.1. Extension of zero-Mach models to the compressible system

We shall develop in parallel our multi-scale turbulence modeling and the related adaptive numerical methods of AeroSol. Without prejudice to methods that will be on the podium in the future, a first step in this direction will be to extend to a compressible framework the continuous temporal hybrid RANS/LES method we have developed up to now in a Mach zero context.

3.2.2.2. Study of wall flows with and without mass or heat transfer at the wall: determination and validation of relevant criteria for hybrid turbulence models

In the targeted application domains, turbulence/wall interactions and heat transfer at the fluid-solid interface are physical phenomena whose numerical prediction is at the heart of the concerns of our industrial partners. For instance, for a jet engine manufacturer, being able to properly design the configuration of the cooling of the walls of its engine combustion chamber in the presence of thermoacoustic instabilities is based on the proper identification and a thorough understanding of the major mechanisms that drive the dynamics of the parietal transfer. Our objective is to take advantage of our analysis, experimental and computational tools to actively participate in the improvement of the collective knowledge of such kind of transfer. The flow configurations dealt with from the beginning of the project are those of subsonic, single-phase impinging jets or JICF (jets in crossflow) with the possible presence of an interacting acoustic wave. The issue of conjugate heat transfer at the wall will be also gradually investigated. The existing switchover criteria of the hybrid RANS/LES models will be tested on these flow configurations in order to determine their domain of validity. In parallel, the hydrodynamic instability modes of the JICF will be studied experimentally and theoretically (in cooperation with the SIAME laboratory) in order to determine the possibility to drive a change of instability regime (e.g., from absolute to convective) and thus to propose challenging flow conditions that would be relevant for the setting-up of an hybrid LES/DNS approach aimed at supplementing the hybrid RANS/LES approach.

3.2.2.3. Improvement of turbulence models

The production and subsequent use of DNS (AeroSol library) and experimental (MAVERIC bench) databases dedicated to the improvement of the physical models is a significant part of our activity. In that respect, our present capability of producing in-situ experimental data for simulation validation and flow analysis is clearly a strongly differentiating mark of our project. The analysis of the DNS and experimental data produced make the improvement of the hybrid RANS/LES approach possible. Our hybrid temporal LES (HTLES) method has a decisive advantage over all other hybrid RANS/LES approaches since it relies on a well-defined time-filtering formalism. This feature greatly facilitates the proper extraction from the databases of the various terms appearing in transport equations obtained at the different scales involved (e.g. from RANS to LES). But we would not be comprehensive in that matter if we were not questioning the relevance of any simulation-experiment comparisons. In other words, a central issue is the following question: are we comparing the same quantities between simulations and experiment? From an experimental point of view, the questions to be raised will be, among others, the possible difference in resolution between the experiment and the simulations, the similar location of the measurement points and simulation points, the acceptable level of random error associated to the necessary finite number of samples. In that respect, the recourse to uncertainty quantification techniques will be advantageously considered.

3.2.3. Development of an efficient implicit high-order compressible solver scalable on new architectures

As the flows simulated are very computationally demanding, we will maintain our efforts in the development of AeroSol in the following directions:

- Efficient implementation of the discontinuous Galerkin method.
- Implicit methods based on Jacobian-Free-Newton-Krylov methods and multigrid.
- Porting on heterogeneous architectures.
- Implementation of models.

3.2.3.1. Efficient implementation of the discontinuous Galerkin method

In high-order discontinuous Galerkin methods, the unknown vector is composed of a concatenation of the unknowns in the cells of the mesh. An explicit residual computation is composed of three loops: an integration loop on the cells, for which computations in two different cells are independent, an integration loop on boundary faces, in which computations depend on data of one cell and on the boundary conditions, and an integration loop on the interior faces, in which computations depend on data of the two neighboring cells. Each of these loops is composed of three steps: the first step consists in interpolating data at the quadrature points; the second step in computing a nonlinear flux at the quadrature points (the physical flux for the cell loop, an upwind flux for interior faces or a flux adapted to the kind of boundary condition for boundary faces); and the third step in projecting the nonlinear flux on the degrees of freedom.

In this research direction, we propose to exploit the strong memory locality of the method (i.e., the fact that all the unknowns of a cell are stocked contiguously). This formulation can reduce the linear steps of the method (interpolation on the quadrature points and projection on the degrees of freedom) to simple matrix-matrix product which can be optimized. For the nonlinear steps, composed of the computation of the physical flux on the cells and of the numerical flux on the faces, we will try to exploit vectorization.

3.2.3.2. Implicit methods based on Jacobian-Free-Newton-Krylov methods and multigrid

For our computations of the IMPACT-AE project, we have used explicit time stepping. The time stepping is limited by the CFL condition, and in our flow, the time step is limited by the acoustic wave velocity. As the Mach number of the flow we simulated in IMPACT-AE was low, the acoustic time restriction is much lower than the turbulent time scale, which is driven by the velocity of the flow. We hope to have a better efficiency by using time implicit methods, for using a time step driven by the velocity of the flow.

Using implicit time stepping in compressible flows in particularly difficult, because the system is fully nonlinear, such that the nonlinear solving theoretically requires to build many times the Jacobian. Our experience in implicit methods is that the building of a Jacobian is very costly, especially in three dimensions and in a high-order framework, because the optimization of the memory usage is very difficult. That is why we propose to use a Jacobian-free implementation, based on [52]. This method consists in solving the linear steps of the Newton method by a Krylov method, which requires Jacobian-vector product. The smart idea of this method is to replace this product by an approximation based on a difference of residual, therefore avoiding any Jacobian computation. Nevertheless, Krylov methods are known to converge slowly, especially for the compressible system when the Mach number is low, because the system is ill-conditioned. In order to precondition, we propose to use an aggregation-based multigrid method, which consists in using the same numerical method on coarser meshes obtained by aggregation of the initial mesh. This choice is driven by the fact that multigrid methods are the only one to scale linearly [61], [62] with the number of unknowns in term of number of operations, and that this preconditioning does not require any Jacobian computation.

Beyond the technical aspects of the multigrid approach, which is challenging to implement, we are also interested in the design of an efficient aggregation. This often means to perform an aggregation based on criteria (anisotropy of the problem, for example) [55]. To this aim, we propose to extend the scalar analysis of [63] to a linearized version of the Euler and Navier-Stokes equations, and try to deduce an optimal strategy for anisotropic aggregation, based on the local characteristics of the flow. Note that discontinuous Galerkin methods are particularly well suited to h-p aggregation, as this kind of methods can be defined on any shape [34].

3.2.3.3. Porting on heterogeneous architectures

Until the beginning of the 2000s, the computing capacities have been improved by interconnecting an increasing number of more and more powerful computing nodes. The computing capacity of each node was increased by improving the clock speed, the number of cores per processor, the introduction of a separate and dedicated memory bus per processor, but also the instruction level parallelism, and the size of the memory cache. Even if the number of transistors kept on growing up, the clock speed improvement has flattened since the mid 2000s [58]. Already in 2003, [49] pointed out the difficulties for efficiently using the biggest clusters: "While these super-clusters have theoretical peak performance in the Teraflops range, sustained performance with real applications is far from the peak. Salinas, one of the 2002 Gordon Bell Awards was able to sustain 1.16 Tflops on ASCI White (less than 10% of peak)." From the current multi-core architectures, the trend is now to use many-core accelerators. The idea behind many-core is to use an accelerator composed of a lot of relatively slow and simplified cores for executing the most simple parts of the algorithm. The larger the part of the code executed on the accelerator, the faster the code may become. Therefore, it is necessary to work on the heterogeneous aspects of computations. These heterogeneities are intrinsic to our computations and have two sources. The first one is the use of hybrid meshes, which are necessary for using a locally-structured mesh in a boundary layer. As the different cell shapes (pyramids, hexahedra, prisms and tetrahedra) do not have the same number of degrees of freedom, nor the same number of quadrature points, the execution time on one face or one cell depends on its shape. The second source of heterogeneity are the boundary conditions. Depending on the kind of boundary conditions, user-defined boundary values might be needed, which induces a different computational cost. Heterogeneities are typically what may decrease efficiency in parallel if the workload is not well balanced between the cores. Note that heterogeneities were not dealt with in what we consider as one of the most advanced work on discontinuous Galerkin on GPU [51], as only straight simplicial cell shapes were addressed. For managing at best our heterogeneous computations on heterogeneous architectures, we propose to use the execution runtime StarPU [33]. For this, the discontinuous Galerkin algorithm will be reformulated in terms of a graph of tasks. The previous tasks on the memory management will be useful for that. The linear steps of the discontinuous Galerkin methods require also memory transfers, and one issue consists in determining the optimal task granularity for this step, i.e. the number of cells or face integrations to be sent in parallel on the accelerator. On top of that, the question of which device is the most appropriate to tackle such kind of tasks is to be discussed.

Last, we point out that the combination of shared-memory and distributed-memory parallel programming models is better suited than only the distributed-memory one for multigrid, because in a hybrid version, a wider part of the mesh shares the same memory, therefore making a coarser aggregation possible.

These aspects will benefit from a particularly stimulating environment in the Inria Bordeaux Sud Ouest center around high-performance computing, which is one of the strategic axes of the center.

3.2.3.4. Implementation of turbulence models in AeroSol and validation

We will gradually insert models developed in research direction 3.2.2.1 in the AeroSol library in which we develop methods for the DNS of compressible turbulent flows at low Mach number. Indeed, due to its formalism based on temporal filtering, the HTLES approach offers a consistent theoretical framework characterized by a continuous transition from RANS to DNS, even for complex flow configurations (e.g. without directions of spatial homogeneity). As for the discontinuous Galerkin method available presently in AeroSol, it is the best suited and versatile method able to meet the requirements of accuracy, stability and cost related to the local (varying) level of resolution of the turbulent flow at hand, regardless of its complexity. The first step in this direction was taken in 2017 during the internship of Axelle Perraud, who has implemented a turbulence model (k-ω-SST) in the Aerosol library.

3.2.4. Validation of the simulations: test flow configurations

To supplement whenever necessary the test flow configuration of MAVERIC and apart from configurations that could emerge in the course of the project, the following configurations for which either experimental data, simulation data or both have been published will be used whenever relevant for benchmarking the quality of our agile computations:

- The impinging turbulent jet (simulations).
- The ORACLES two-channel dump combustor developed in the European projects LES4LPP and MOLECULES.
- The non reactive single-phase PRECCINSTA burner (monophasic swirler), a configuration that has been extensively calculated in particular with the AVBP and Yales2 codes.
- The LEMCOTEC configuration (monophasic swirler + effusion cooling).
- The ONERA MERCATO two-phase injector configuration provided the question of confidentiality of the data is not an obstacle.
- Rotating turbulent flows with wall interaction and heat transfer.
- Turbulent flows with buoyancy.

CARDAMOM Project-Team

3. Research Program

3.1. Variational discrete asymptotic modelling

In many of the applications we consider, intermediate fidelity models are or can be derived using an asymptotic expansion for the relevant scale resolving PDEs, and eventually considering some averaged for of the resulting continuous equations. The resulting systems of PDEs are often very complex and their characterization, e.g. in terms of stability, unclear, or poor, or too complex to allow to obtain discrete analogy of the continuous properties. This makes the numerical approximation of these PDE systems a real challenge. Moreover, most of these models are often based on asymptotic expansions involving small geometrical scales. This is true for many applications considered here involving flows in/of thin layers (free surface waves, liquid films on wings generating ice layers, oxide flows in material cracks, etc). This asymptotic expansion is nothing else than a discretization (some sort of Taylor expansion) in terms of the small parameter. The actual discretization of the PDE system is another expansion in space involving as a small parameter the mesh size. What is the interaction between these two expansions ? Could we use the spatial discretization (truncation error) as means of filtering undesired small scales instead of having to explicitly derive PDEs for the large scales ? We will investigate in depth the relations between asymptotics and discretization by :

- comparing the asymptotic limits of discretized forms of the relevant scale resolving equations with the discretization of the analogous continuous asymptotic PDEs. Can we discretize a well understood system of PDEs instead of a less understood and more complex one ? ;
- study the asymptotic behaviour of error terms generated by coarse one-dimensional discretization in the direction of the “small scale”. What is the influence of the number of cells along the vertical direction, and of their clustering ? ;
- derive equivalent continuous equations (modified equations) for anisotropic discretizations in which the direction is direction of the “small scale” is approximated with a small number of cells. What is the relation with known asymptotic PDE systems ?

Our objective is to gain sufficient control of the interaction between discretization and asymptotics to be able to replace the coupling of several complex PDE systems by adaptive strongly anisotrotropic finite elemient approximations of relevant and well understood PDEs. Here the anisotropy is intended in the sense of having a specific direction in which a much poorer (and possibly variable with the flow conditions) polynomial approximation (expansion) is used. The final goal is, profiting from the availability of faster and cheaper computational platforms, to be able to automatically control numerical *and* physical accuracy of the model with the same techniques. This activity will be used to improve our modelling in coastal engineering as well as for de-anti icing systems, wave energy converters, composite materials (cf. next sections).

In parallel to these developments, we will make an effort in to gain a better understanding of continuous asymptotic PDE models. We will in particular work on improving, and possibly, simplifying their numerical approximation. An effort will be done in trying to embed in these more complex nonlinear PDE models discrete analogs of operator identities necessary for stability (see e.g. the recent work of [70], [72] and references therein).

3.2. High order discretizations on moving adaptive meshes

We will work on both the improvement of high order mesh generation and adaptation techniques, and the construction of more efficient, adaptive high order discretisation methods.

Concerning curved mesh generation, we will focus on two points. First propose a robust and automatic method to generate curved simplicial meshes for realistic geometries. The untangling algorithm we plan to develop is a hybrid technique that gathers a local mesh optimization applied on the surface of the domain and a linear elasticity analogy applied in its volume. Second we plan to extend the method proposed in [26] to hybrid meshes (prism/tetra).

For time dependent adaptation we will try to exploit as much as possible the use of $r-$adaptation techniques based on the solution of some PDE system for the mesh. We will work on enhancing the results of [29] by developing more robust nonlinear variants allowing to embed rapidly moving objects. For this the use of non-linear mesh PDEs (cf e.g. [80], [85], [38]), combined with Bezier type approximations for the mesh displacements to accommodate high order curved meshes [26], and with improved algorithms to discretize accurately and fast the elliptic equations involved. For this we will explore different type of relaxation methods, including those proposed in [71], [75], [74] allowing to re-use high order discretizations techniques already used for the flow variables. All these modelling approaches for the mesh movement are based on some minimization argument, and do not allow easily to take into account explicitly properties such as e.g. the positivity of nodal volumes. An effort will be made to try to embed these properties, as well as to improve the control on the local mesh sizes obtained. Developments made in numerical methods for Lagrangian hydrodynamics and compressible materials may be a possible path for these objectives (see e.g. [49], [91], [90] and references therein). We will stretch the use of these techniques as much as we can, and couple them with remeshing algorithms based on local modifications plus conservative, high order, and monotone ALE (or other) remaps (cf. [27], [58], [92], [47] and references therein).
The development of high order schemes for the discretization of the PDE will be a major part of our activity. We will work from the start in an Arbitrary Lagrangian Eulerian setting, so that mesh movement will be easily accommodated, and investigate the following main points:

- the ALE formulation is well adapted both to handle moving meshes, and to provide conservative, high order, and monotone remaps between different meshes. We want to address the issue of cost-accuracy of adaptive mesh computations by exploring different degrees of coupling between the flow and the mesh PDEs. Initial experience has indicated that a clever coupling may lead to a considerable CPU time reduction for a given resolution [29]. This balance is certainly dependent on the nature of the PDEs, on the accuracy level sought, on the cost of the scheme, and on the time stepping technique. All these elements will be taken into account to try to provide the most efficient formulation ;
- the conservation of volume, and the subsequent preservation of constant mass-momentum-energy states on deforming domains is one of the most primordial elements of Arbitrary Lagrangian-Eulerian formulations. For complex PDEs as the ones considered here, of especially for some applications, there may be a competition between the conservation of e.g. mass, an the conservation of other constant states, as important as mass. This is typically the case for free surface flows, in which mass preservation is in competitions with the preservation of constant free surface levels [29]. Similar problems may arise in other applications. Possible solutions to this competition may come from super-approximation (use of higher order polynomials) of some of the data allowing to reduce (e.g. bathymetry) the error in the preservation of one of the competing quantities. This is similar to what is done in super-parametric approximations of the boundaries of an object immersed in the flow, except that in our case the data may enter the PDE explicitly and not only through the boundary conditions. Several efficient solutions for this issue will be investigated to obtain fully conservative moving mesh approaches:
- an issue related to the previous one is the accurate treatment of wall boundaries. It is known that even for standard lower order (second) methods, a higher order, curved, approximation of the boundaries may be beneficial. This, however, may become difficult when considering moving objects, as in the case e.g. of the study of the impact of ice debris in the flow. To alleviate this issue, we plan to follow on with our initial work on the combined use of immersed boundaries techniques with high order, anisotropic (curved) mesh adaptation. In particular, we will develop combined approaches involving high order hybrid meshes on fixed boundaries with the use of penalization techniques and immersed

boundaries for moving objects. We plan to study the accuracy obtainable across discontinuous functions with $r-$adaptive techniques, and otherwise use whenever necessary anisotropic meshes to be able to provide a simplified high order description of the wall boundary (cf. [69]). The use of penalization will also provide a natural setting to compute immediate approximations of the forces on the immersed body [73], [76]. An effort will be also made on improving the accuracy of these techniques using e.g. higher order approaches, either based on generalizations of classical splitting methods [59], or on some iterative Defect Correction method (see e.g. [40]) ;

- the proper treatment of different physics may be addressed by using mixed/hybrid schemes in which different variables/equations are approximated using a different polynomial expansion. A typical example is our work on the discretization of highly non-linear wave models [54] in which we have shown how to use a standard continuous Galerkin method for the elliptic equation/variable representative of the dispersive effects, while the underlying hyperbolic system is evolved using a (discontinuous) third order finite volume method. This technique will be generalized to other classes of discontinuous methods, and similar ideas will be used in other context to provide a flexible approximation. Such mathods have clear advantages in multiphase flows but not only. A typical example where such mixed methods are beneficial are flows involving different species and tracer equations, which are typically better treated with a discontinuous approximation. Another example is the use of this mixed approximation to describe the topography with a high order continuous polynomial even in discontinuous method. This allows to greatly simplify the numerical treatment of the bathymetric source terms ;
- the enhancement of stabilized methods based on some continuous finite element approximation will remain a main topic. We will further pursue the study on the construction of simplified stabilization operators which do not involve any contributions to the mass matrix. We will in particular generalize our initial results to higher order spatial approximations using cubature points, or Bezier polynomials, or also hierarchical approximations. This will also be combined with time dependent variants of the reconstruction techniques initially proposed by D. Caraeni [39], allowing to have a more flexible approach similar to the so-called P^nP^m method [52], [84]. How to localize these enhancements, and to efficiently perform local reconstructions/enrichment, as well as $p-$adaptation, and handling hanging nodes will also be a main line of work. A clever combination of hierarchical enrichment of the polynomials, with a constrained approximation will be investigated. All these developments will be combined with the shock capturing/positivity preserving construction we developed in the past. Other discontinuity resolving techniques will be investigated as well, such as face limiting techniques as those partially studied in [56] ;
- time stepping is an important issue, especially in presence of local mesh adaptation. The techniques we use will force us to investigate local and multilevel techniques. We will study the possibility constructing semi-implicit methods combining extrapolation techniques with space-time variational approaches. Other techniques will be considered, as multi-stage type methods obtained using Defect-Correction, Multi-step Runge-Kutta methods [36], as well as spatial partitioning techniques [65]. A major challenge will be to be able to guarantee sufficient locality to the time integration method to allow to efficiently treat highly refined meshes, especially for viscous reactive flows. Another challenge will be to embed these methods in the stabilized methods we will develop.

3.3. Coupled approximation/adaptation in parameter and physical space

As already remarked, classical methods for uncertainty quantification are affected by the so-called Curse-of-Dimensionality. Adaptive approaches proposed so far, are limited in terms of efficiency, or of accuracy. Our aim here is to develop methods and algorithms permitting a very high-fidelity simulation in the physical and in the stochastic space at the same time. We will focus on both non-intrusive and intrusive approaches.

Simple non-intrusive techniques to reduce the overall cost of simulations under uncertainty will be based on adaptive quadrature in stochastic space with mesh adaptation in physical space using error monitors related to the variance of to the sensitivities obtained e.g. by an ANOVA decomposition. For steady state problems,

remeshing using metric techniques is enough. For time dependent problems both mesh deformation and remeshing techniques will be used. This approach may be easily used in multiple space dimensions to minimize the overall cost of model evaluations by using high order moments of the properly chosen output functional for the adaptation (as in optimization). Also, for high order curved meshes, the use of high order moments and sensitivities issued from the UQ method or optimization provides a viable solution to the lack of error estimators for high order schemes.

Despite the coupling between stochastic and physical space, this approach can be made massively parallel by means of extrapolation/interpolation techniques for the high order moments, in time and on a reference mesh, guaranteeing the complete independence of deterministic simulations. This approach has the additional advantage of being feasible for several different application codes due to its non-intrusive character.
To improve on the accuracy of the above methods, intrusive approaches will also be studied. To propagate uncertainties in stochastic differential equations, we will use Harten's multiresolution framework, following [25]. This framework allows a reduction of the dimensionality of the discrete space of function representation, defined in a proper stochastic space. This reduction allows a reduction of the number of explicit evaluations required to represent the function, and thus a gain in efficiency. Moreover, multiresolution analysis offers a natural tool to investigate the local regularity of a function and can be employed to build an efficient refinement strategy, and also provides a procedure to refine/coarsen the stochastic space for unsteady problems. This strategy should allow to capture and follow all types of flow structures, and, as proposed in [25], allows to formulate a non-linear scheme in terms of compression capabilities, which should allow to handle non-smooth problems. The potential of the method also relies on its moderate intrusive behaviour, compared to e.g. spectral Galerkin projection, where a theoretical manipulation of the original system is needed.
Several activities are planned to generalize our initial work, and to apply it to complex flows in multiple (space) dimensions and with many uncertain parameters.

The first is the improvement of the efficiency. This may be achieved by means of anisotropic mesh refinement, and by experimenting with a strong parallelization of the method. Concerning the first point, we will investigate several anisotropic refinement criteria existing in literature (also in the UQ framework), starting with those already used in the team to adapt the physical grid. Concerning the implementation, the scheme formulated in [25] is conceived to be highly parallel due to the external cycle on the number of dimensions in the space of uncertain parameters. In principle, a number of parallel threads equal to the number of spatial cells could be employed. The scheme should be developed and tested for treating unsteady and discontinuous probability density function, and correlated random variables. Both the compression capabilities and the accuracy of the scheme (in the stochastic space) should be enhanced with a high-order multidimensional conservative and non-oscillatory polynomial reconstruction (ENO/WENO).

Another main objective is related to the use of multiresolution in both physical and stochastic space. This requires a careful handling of data and an updated definition of the wavelet. Until now, only a weak coupling has been performed, since the number of points in the stochastic space varies according to the physical space, but the number of points in the physical space remains unchanged. Several works exist on the multiresolution approach for image compression, but this could be the first time i in which this kind of approach would be applied at the same time in the two spaces with an unsteady procedure for refinement (and coarsening). The experimental code developed using these technologies will have to fully exploit the processing capabilities of modern massively parallel architectures, since there is a unique mesh to handle in the coupled physical/stochastic space.

3.4. Robust multi-fidelity modelling for optimization and certification

Due to the computational cost, it is of prominent importance to consider multi-fidelity approaches gathering high-fidelity and low-fidelity computations. Note that low-fidelity solutions can be given by both the use of surrogate models in the stochastic space, and/or eventually some simplified choices of physical models of some element of the system. Procedures which deal with optimization considering uncertainties for complex problems may require the evaluation of costly objective and constraint functions hundreds or even thousands of times. The associated costs are usually prohibitive. For these reason, the robustness of the optimal

solution should be assessed, thus requiring the formulation of efficient methods for coupling optimization and stochastic spaces. Different approaches will be explored. Work will be developed along three axes:

1. a robust strategy using the statistics evaluation will be applied separately, *i.e.* using only low or high-fidelity evaluations. Some classical optimization algorithms will be used in this case. Influence of high-order statistics and model reduction in the robust design optimization will be explored, also by further developing some low-cost methods for robust design optimization working on the so-called Simplex^2 method [45] ;
2. a multi-fidelity strategy by using in an efficient way low fidelity and high-fidelity estimators both in physical and stochastic space will be conceived, by using a Bayesian framework for taking into account model discrepancy and a PC expansion model for building a surrogate model ;
3. develop advanced methods for robust optimization. In particular, the Simplex^2 method will be modified for introducing a hierarchical refinement with the aim to reduce the number of stochastic samples according to a given design in an adaptive way.

This work is related to the activities foreseen in the EU contract MIDWEST, in the ANR LabCom project VIPER (currently under evaluation), in a joint project with DGA and VKI, in two projects under way with AIRBUS and SAFRAN-HERAKLES.

CELESTE Project-Team

3. Research Program

3.1. General presentation

Our objectives correspond to four major challenges of machine learning where mathematical statistics have a key role. First, any machine learning procedure depends on hyperparameters that must be chosen, and many procedures are available for any given learning problem: both are an estimator selection problem. Second, with high-dimensional and/or large data, the computational complexity of algorithms must be taken into account differently, leading to possible trade-offs between statistical accuracy and complexity, for machine learning procedures themselves as well as for estimator selection procedures. Third, real data are almost always corrupted partially, making it necessary to provide learning (and estimator selection) procedures that are robust to outliers and heavy tails, while being able to handle large datasets. Fourth, science currently faces a reproducibility crisis, making it necessary to provide statistical inference tools (p-values, confidence regions) for assessing the significance of the output of any learning algorithm (including the tuning of its hyperparameters), in a computationally efficient way.

3.2. Estimator selection

An important goal of CELESTE is to build and study procedures that can deal with general estimators (especially those actually used in practice, which often rely on some optimization algorithm), such as cross-validation and Lepski's method. In order to be practical, estimator selection procedures must be fully data-driven (that is, not relying on any unknown quantity), computationally tractable (especially in the high-dimensional setting, for which specific procedures must be developed) and robust to outliers (since most real data sets include a few outliers). CELESTE aims at providing a precise theoretical analysis (for new and existing popular estimator selection procedures), that explains as well as possible their observed behaviour in practice.

3.3. Relating statistical accuracy to computational complexity

When several learning algorithms are available, with increasing computational complexity and statistical performance, which one should be used, given the amount of data and the computational power available? This problem has emerged as a key question induced by the challenge of analyzing large amounts of data – the "big data" challenge. CELESTE wants to tackle the major challenge of understanding the time-accuracy trade-off, which requires providing new statistical analyses of machine learning procedures – as they are done in practice, including optimization algorithms – that are *precise enough* in order to account for differences of performance observed in practice, leading to general conclusions that can be trusted more generally. For instance, we study the performance of ensemble methods combined with subsampling, which is a common strategy for handling big data; examples include random forests and median-of-means algorithms.

3.4. Robustness to outliers and heavy tails (with tractable algorithms)

The classical theory of robustness in statistics has recently received a lot of attention in the machine learning community. The reason is simple: large datasets are easily corrupted, due to – for instance – storage and transmission issues, and most learning algorithms are highly sensitive to dataset corruption. For example, the lasso can be completely misled by the presence of even a single outlier in a dataset. A major challenge in robust learning is to provide computationally tractable estimators with optimal subgaussian guarantees. A second important challenge in robust learning is to deal with datasets where every (x_i, y_i) is slightly corrupted. In large-dimensional data, every single data point x_i is likely to have several corrupted coordinates, and no estimator currently has strong theoretical guarantees for such data. A third important challenge is that of

robust estimator selection or aggregation. Even if several robust estimators can be built, the final aggregation or selection step in a user's routine is usually based on empirical means. This is not robust, and may damage the global performance of the procedure. Instead, we can consider more sophisticated types of aggregation of the base robust estimators built so far. A convenient framework to do so is called adversarial learning (also known as: prediction of individual sequences). Here, data is not assumed to be stochastic, and it could even be chosen by an adversary.

3.5. Statistical inference: (multiple) tests and confidence regions (including post-selection)

CELESTE considers the problems of quantifying the uncertainty of predictions or estimations (thanks to confidence intervals) and of providing significance levels (p-values, corrected for multiplicity if needed) for each "discovery" made by a learning algorithm. This is an important practical issue when performing feature selection – one then speaks of post-selection inference – change-point detection or outlier detection, to name but a few. We tackle it in particular through a collaboration with the Parietal team (Inria Saclay) and LBBE (CNRS), with applications in neuroimaging and genomics.

COMMANDS Project-Team

3. Research Program

3.1. Historical aspects

The roots of deterministic optimal control are the "classical" theory of the calculus of variations, illustrated by the work of Newton, Bernoulli, Euler, and Lagrange (whose famous multipliers were introduced in [24]), with improvements due to the "Chicago school", Bliss [16] during the first part of the 20th century, and by the notion of relaxed problem and generalized solution (Young [29]).

Trajectory optimization really started with the spectacular achievement done by Pontryagin's group [28] during the fifties, by stating, for general optimal control problems, nonlocal optimality conditions generalizing those of Weierstrass. This motivated the application to many industrial problems (see the classical books by Bryson and Ho [20], Leitmann [26], Lee and Markus [25], Ioffe and Tihomirov [23]).

Dynamic programming was introduced and systematically studied by R. Bellman during the fifties. The HJB equation, whose solution is the value function of the (parameterized) optimal control problem, is a variant of the classical Hamilton-Jacobi equation of mechanics for the case of dynamics parameterized by a control variable. It may be viewed as a differential form of the dynamic programming principle. This nonlinear first-order PDE appears to be well-posed in the framework of *viscosity solutions* introduced by Crandall and Lions [21]. The theoretical contributions in this direction did not cease growing, see the books by Barles [14] and Bardi and Capuzzo-Dolcetta [13].

3.2. Trajectory optimization

The so-called *direct methods* consist in an optimization of the trajectory, after having discretized time, by a nonlinear programming solver that possibly takes into account the dynamic structure. So the two main problems are the choice of the discretization and the nonlinear programming algorithm. A third problem is the possibility of refinement of the discretization once after solving on a coarser grid.

In the *full discretization approach*, general Runge-Kutta schemes with different values of control for each inner step are used. This allows to obtain and control high orders of precision, see Hager [22], Bonnans [17]. In the *indirect* approach, the control is eliminated thanks to Pontryagin's maximum principle. One has then to solve the two-points boundary value problem (with differential variables state and costate) by a single or multiple shooting method. The questions are here the choice of a discretization scheme for the integration of the boundary value problem, of a (possibly globalized) Newton type algorithm for solving the resulting finite dimensional problem in $I\!R^n$ (n is the number of state variables), and a methodology for finding an initial point.

3.3. Hamilton-Jacobi-Bellman approach

This approach consists in calculating the value function associated with the optimal control problem, and then synthesizing the feedback control and the optimal trajectory using Pontryagin's principle. The method has the great particular advantage of reaching directly the global optimum, which can be very interesting when the problem is not convex.

Optimal stochastic control problems occur when the dynamical system is uncertain. A decision typically has to be taken at each time, while realizations of future events are unknown (but some information is given on their distribution of probabilities). In particular, problems of economic nature deal with large uncertainties (on prices, production and demand). Specific examples are the portfolio selection problems in a market with risky and non-risky assets, super-replication with uncertain volatility, management of power resources (dams, gas). Air traffic control is another example of such problems.

For solving stochastic control problems, we studied the so-called Generalized Finite Differences (GFD), that allow to choose at any node, the stencil approximating the diffusion matrix up to a certain threshold [19]. Determining the stencil and the associated coefficients boils down to a quadratic program to be solved at each point of the grid, and for each control. This is definitely expensive, with the exception of special structures where the coefficients can be computed at low cost. For two dimensional systems, we designed a (very) fast algorithm for computing the coefficients of the GFD scheme, based on the Stern-Brocot tree [18].

CQFD Project-Team

3. Research Program

3.1. Introduction

The scientific objectives of the team are to provide mathematical tools for modeling and optimization of complex systems. These systems require mathematical representations which are in essence dynamic, multi-model and stochastic. This increasing complexity poses genuine scientific challenges in the domain of modeling and optimization. More precisely, our research activities are focused on stochastic optimization and (parametric, semi-parametric, multidimensional) statistics which are complementary and interlinked topics. It is essential to develop simultaneously statistical methods for the estimation and control methods for the optimization of the models.

3.2. Main research topics

Stochastic modeling: Markov chain, Piecewise Deterministic Markov Processes (PDMP), Markov Decision Processes (MDP).
The mathematical representation of complex systems is a preliminary step to our final goal corresponding to the optimization of its performance. The team CQFD focuses on two complementary types of approaches. The first approach is based on mathematical representations built upon physical models where the dynamic of the real system is described by *stochastic processes*. The second one consists in studying the modeling issue in an abstract framework where the real system is considered as black-box. In this context, the outputs of the system are related to its inputs through a *statistical model*. Regarding stochastic processes, the team studies Piecewise Deterministic Markov Processes (PDMPs) and Markov Decision Processes (MDPs). These two classes of Markov processes form general families of controlled stochastic models suitable for the design of sequential decision-making problems. They appear in many fields such as biology, engineering, computer science, economics, operations research and provide powerful classes of processes for the modeling of complex systems. Our contribution to this topic consists in expressing real-life industrial problems into these mathematical frameworks. Regarding statistical methods, the team works on dimension reduction models. They provide a way to understand and visualize the structure of complex data sets. Furthermore, they are important tools in several different areas such as data analysis and machine learning, and appear in many applications such as biology, genetics, environment and recommendation systems. Our contribution to this topic consists in studying semiparametric modeling which combines the advantages of parametric and nonparametric models.

Estimation methods: estimation for PDMP; estimation in non- and semi- parametric regression modeling.
To the best of our knowledge, there does not exist any general theory for the problems of estimating parameters of PDMPs although there already exist a large number of tools for sub-classes of PDMPs such as point processes and marked point processes. To fill the gap between these specific models and the general class of PDMPs, new theoretical and mathematical developments will be on the agenda of the whole team. In the framework of non-parametric regression or quantile regression, we focus on kernel estimators or kernel local linear estimators for complete data or censored data. New strategies for estimating semi-parametric models via recursive estimation procedures have also received an increasing interest recently. The advantage of the recursive estimation approach is to take into account the successive arrivals of the information and to refine, step after step, the implemented estimation algorithms. These recursive methods do require restarting calculation of parameter estimation from scratch when new data are added to the base. The idea is to use only the previous estimations and the new data to refresh the estimation. The gain in time could be very interesting and there are many applications of such approaches.

Dimension reduction: dimension-reduction via SIR and related methods, dimension-reduction via multidimensional and classification methods.

Most of the dimension reduction approaches seek for lower dimensional subspaces minimizing the loss of some statistical information. This can be achieved in modeling framework or in exploratory data analysis context.
In modeling framework we focus our attention on semi-parametric models in order to conjugate the advantages of parametric and nonparametric modeling. On the one hand, the parametric part of the model allows a suitable interpretation for the user. On the other hand, the functional part of the model offers a lot of flexibility. In this project, we are especially interested in the semi-parametric regression model $Y = f(X'\theta) + \varepsilon$, the unknown parameter θ belongs to $\mathbb{R}^p$ for a single index model, or is such that $\theta = [\theta_1, \cdots, \theta_d]$ (where each θ_k belongs to $\mathbb{R}^p$ and $d \leq p$ for a multiple indices model), the noise ε is a random error with unknown distribution, and the link function f is an unknown real valued function. Another way to see this model is the following: the variables X and Y are independent given $X'\theta$. In our semi-parametric framework, the main objectives are to estimate the parametric part θ as well as the nonparametric part which can be the link function f, the conditional distribution function of Y given X or the conditional quantile q_α. In order to estimate the dimension reduction parameter θ we focus on the Sliced Inverse Regression (SIR) method which has been introduced by Li [37] and Duan and Li [35].
Methods of dimension reduction are also important tools in the field of data analysis, data mining and machine learning.They provide a way to understand and visualize the structure of complex data sets.Traditional methods among others are principal component analysis for quantitative variables or multiple component analysis for qualitative variables. New techniques have also been proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few observation units. In this context, we focus on the problem of synthetic variables construction, whose goals include increasing the predictor performance and building more compact variables subsets. Clustering of variables is used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. For a review, see, e.g., the textbook of Duda [36].

Stochastic control: optimal stopping, impulse control, continuous control, linear programming.
The main objective is to develop *approximation techniques* to provide quasi-optimal feasible solutions and to derive *optimality results* for control problems related to MDPs and PDMPs:

- *Approximation techniques*. The analysis and the resolution of such decision models mainly rely on the maximum principle and/or the dynamic/linear programming techniques together with their various extensions such as the value iteration (VIA) and the policy iteration (PIA) algorithm. However, it is well known that these approaches are hardly applicable in practice and suffer from the so-called *curse of dimensionality*. Hence, solving numerically a PDMP or an MDP is a difficult and important challenge. Our goal is to obtain results which are both consistent from a theoretical point of view and computationally tractable and accurate from an application standpoint. It is important to emphasize that these research objectives were not planned in our initial 2009 program.

  Our objective is to propose approximation techniques to efficiently compute the optimal value function and to get quasi-optimal controls for different classes of constrained and unconstrained MDPs with general state/action spaces, and possibly unbounded cost function. Our approach is based on combining the linear programming formulation of an MDP with probabilistic approximation techniques related to quantization techniques and the theory of empirical processes. An other aim is to apply our methods to specific industrial applications in collaboration with industrial partners such as Airbus Defence & Space, Naval Group and Thales.

  Asymptotic approximations are also developed in the context of queueing networks, a class of models where the decision policy of the underlying MDP is in some sense fixed a priori, and our main goal is to study the transient or stationary behavior of the induced Markov process. Even though the decision policy is fixed, these models usually remain intractable to solve. Given this complexity, the team has developed analyses in some limiting regime of practical interest, i.e., queueing models in the large-network, heavy-traffic, fluid or mean-field limit. This approach is helpful to obtain a simpler mathematical description of the system under investigation, which is often given in terms of ordinary differential equations or convex optimization problems.

- *Optimality results*. Our aim is to investigate new important classes of optimal stochastic control problems including constraints and combining continuous and impulse actions for MDPs and PDMPs. In this framework, our objective is to obtain different types of optimality results. For example, we intend to provide conditions to guarantee the existence and uniqueness of the optimality equation for the problem under consideration and to ensure existence of an optimal (and ϵ-optimal) control strategy. We also plan to analyze the structural properties of the optimal strategies as well as to study the associated infinite dimensional linear programming problem. These results can be seen as a first step toward the development of numerical approximation techniques in the sense described above.

DEFI Project-Team

3. Research Program

3.1. Research Program

The research activity of our team is dedicated to the design, analysis and implementation of efficient numerical methods to solve inverse and shape/topological optimization problems, eventually including system uncertainties, in connection with wave imaging, structural design, non-destructive testing and medical imaging modalities. We are particularly interested in the development of fast methods that are suited for real-time applications and/or large scale problems. These goals require to work on both the physical and the mathematical models involved and indeed a solid expertise in related numerical algorithms. A part of the research activity is also devoted to take into account system uncertainties in the solving of inverse/optimization problems. At the interface of physics, mathematics, and computer science, Uncertainty Quantification (UQ) focuses on the development of frameworks and methods to characterize uncertainties in predictive computations. Uncertainties and errors arise at different stages of the numerical simulation. First, errors are introduced due to the physical simplifications in the mathematical modeling of the system investigated; other errors come from the numerical resolution of the mathematical model, due in particular to finite discretization and computations with finite accuracy and tolerance; finally, errors are due a limited knowledge of input quantities (parameters) appearing in the definition of the numerical model being solved.

This section intends to give a general overview of our research interests and themes. We choose to present them through the specific academic example of inverse scattering problems (from inhomogeneities), which is representative of foreseen developments on both inversion and (topological) optimization methods. The practical problem would be to identify an inclusion from measurements of diffracted waves that result from the interaction of the sought inclusion with some (incident) waves sent into the probed medium. Typical applications include biomedical imaging where using micro-waves one would like to probe the presence of pathological cells, or imaging of urban infrastructures where using ground penetrating radars (GPR) one is interested in finding the location of buried facilities such as pipelines or waste deposits. This kind of applications requires in particular fast and reliable algorithms.

By “imaging” we refer to the inverse problem where the concern is only the location and the shape of the inclusion, while “identification” may also indicate getting informations on the inclusion physical parameters.

Both problems (imaging and identification) are non linear and ill-posed (lack of stability with respect to measurements errors if some careful constrains are not added). Moreover, the unique determination of the geometry or the coefficients is not guaranteed in general if sufficient measurements are not available. As an example, in the case of anisotropic inclusions, one can show that an appropriate set of data uniquely determine the geometry but not the material properties.

These theoretical considerations (uniqueness, stability) are not only important in understanding the mathematical properties of the inverse problem, but also guide the choice of appropriate numerical strategies (which information can be stably reconstructed) and also the design of appropriate regularization techniques. Moreover, uniqueness proofs are in general constructive proofs, i.e. they implicitly contain a numerical algorithm to solve the inverse problem, hence their importance for practical applications. The sampling methods introduced below are one example of such algorithms.

A large part of our research activity is dedicated to numerical methods applied to the first type of inverse problems, where only the geometrical information is sought. In its general setting the inverse problem is very challenging and no method can provide universally satisfying solution (respecting the balance cost-precision-stability). This is why in the majority of the practically employed algorithms, some simplification of the underlying mathematical model is used, according to the specific configuration of the imaging experiment. The most popular ones are geometric optics (the Kirchhoff approximation) for high frequencies and weak scattering (the Born approximation) for small contrasts or small obstacles. They actually give full satisfaction

for a wide range of applications as attested by the large success of existing imaging devices (radar, sonar, ultrasound, X-ray tomography, etc.), that rely on one of these approximations.

In most cases, the used simplification result in a linearization of the inverse problem and therefore is usually valid only if the latter is weakly non-linear. The development of simplified models and the improvement of their efficiency is still a very active research area. With that perspective, we are particularly interested in deriving and studying higher order asymptotic models associated with small geometrical parameters such as: small obstacles, thin coatings, wires, periodic media, ... Higher order models usually introduce some non linearity in the inverse problem, but are in principle easier to handle from the numerical point of view than in the case of the exact model.

A larger part of our research activity is dedicated to algorithms that avoid the use of such approximations and that are efficient where classical approaches fail: i.e. roughly speaking when the non linearity of the inverse problem is sufficiently strong. This type of configuration is motivated by the applications mentioned below, and occurs as soon as the geometry of the unknown media generates non negligible multiple scattering effects (multiply-connected and closely spaces obstacles) or when the used frequency is in the so-called resonant region (wave-length comparable to the size of the sought medium). It is therefore much more difficult to deal with and requires new approaches. Our ideas to tackle this problem is mainly motivated and inspired by recent advances in shape and topological optimization methods and in so-called sampling methods.

Sampling methods are fast imaging solvers adapted to multi-static data (multiple receiver-transmitter pairs) at a fixed frequency. Even if they do not use any linearization the forward model, they rely on computing the solutions to a set of linear problems of small size, that can be performed in a completely parallel procedure. Our team has already a solid expertise in these methods applied to electromagnetic 3-D problems. The success of such approaches was their ability to provide a relatively quick algorithm for solving 3-D problems without any need for a priori knowledge on the physical parameters of the targets. These algorithms solve only the imaging problem, in the sense that only the geometrical information is provided.

Despite the large efforts already spent in the development of this type of methods, either from the algorithmic point of view or the theoretical one, numerous questions are still open. These attractive new algorithms also suffer from the lack of experimental validations, due to their relatively recent introduction. We also would like to invest on this side by developing collaborations with engineering research groups that have experimental facilities. From the practical point of view, the most potential limitation of sampling methods would be the need of a large amount of data to achieve a reasonable accuracy. On the other hand, optimization methods do not suffer from this constrain but they require good initial guess to ensure convergence and reduce the number of iterations. Therefore it seems natural to try to combine the two class of methods in order to calibrate the balance between cost and precision.

Among various shape optimization methods, the Level Set method seems to be particularly suited for such a coupling. First, because it shares similar mechanism as sampling methods: the geometry is captured as a level set of an “indicator function” computed on a cartesian grid. Second, because the two methods do not require any a priori knowledge on the topology of the sought geometry. Beyond the choice of a particular method, the main question would be to define in which way the coupling can be achieved. Obvious strategies consist in using one method to pre-process (initialization) or post-process (find the level set) the other. But one can also think of more elaborate ones, where for instance a sampling method can be used to optimize the choice of the incident wave at each iteration step. The latter point is closely related to the design of so called “focusing incident waves” (which are for instance the basis of applications of the time-reversal principle). In the frequency regime, these incident waves can be constructed from the eigenvalue decomposition of the data operator used by sampling methods. The theoretical and numerical investigations of these aspects are still not completely understood for electromagnetic or elastodynamic problems.

Other topological optimization methods, like the homogenization method or the topological gradient method, can also be used, each one provides particular advantages in specific configurations. It is evident that the development of these methods is very suited to inverse problems and provide substantial advantage compared to classical shape optimization methods based on boundary variation. Their applications to inverse problems has not been fully investigated. The efficiency of these optimization methods can also be increased for adequate

asymptotic configurations. For instance small amplitude homogenization method can be used as an efficient relaxation method for the inverse problem in the presence of small contrasts. On the other hand, the topological gradient method has shown to perform well in localizing small inclusions with only one iteration.

A broader perspective would be the extension of the above mentioned techniques to time-dependent cases. Taking into account data in time domain is important for many practical applications, such as imaging in cluttered media, the design of absorbing coatings or also crash worthiness in the case of structural design.

For the identification problem, one would like to also have information on the physical properties of the targets. Of course optimization methods is a tool of choice for these problems. However, in some applications only a qualitative information is needed and obtaining it in a cheaper way can be performed using asymptotic theories combined with sampling methods. We also refer here to the use of so called transmission eigenvalues as qualitative indicators for non destructive testing of dielectrics.

We are also interested in parameter identification problems arising in diffusion-type problems. Our research here is mostly motivated by applications to the imaging of biological tissues with the technique of Diffusion Magnetic Resonance Imaging (DMRI). Roughly speaking DMRI gives a measure of the average distance travelled by water molecules in a certain medium and can give useful information on cellular structure and structural change when the medium is biological tissue. In particular, we would like to infer from DMRI measurements changes in the cellular volume fraction occurring upon various physiological or pathological conditions as well as the average cell size in the case of tumor imaging. The main challenges here are 1) correctly model measured signals using diffusive-type time-dependent PDEs 2) numerically handle the complexity of the tissues 3) use the first two to identify physically relevant parameters from measurements. For the last point we are particularly interested in constructing reduced models of the multiple-compartment Bloch-Torrey partial differential equation using homogenization methods.

The Team devotes a large effort focused on the formulation, implementation and validation of numerical methods for using scientific computing to drive experiments and available data (coming from models, simulation and experiments) by taking into account the system uncertainty. The team is also invested in exploiting the intimate relationship between optimisation and UQ to make Optimisation Under Uncertainty (OUU) tractable. A part of these activities is declined to the simulation of high-fidelity models for fluids, in three main fields, aerospace, energy and environment.

The Team is working on developing original UQ representations and algorithms to deal with complex and large scale models, having high dimensional input parameters with complexes influences. We are organizing our core research activities along different methodological UQ developments related to the challenges discussed above. Obviously, some efforts are shared by different initiatives or projects, and some of them include the continuous improvement of the non-intrusive methods constituting our software libraries. These actions are not detailed in the following, to focus the presentation on more innovative aspects, but we mentioned nonetheless the continuous developments and incorporation in our libraries of advanced sparse grid methods, sparsity promoting strategies and low rank methods.

An effort is dedicated to the efficient construction of surrogate models that are central in both forward and backward UQ problems, aiming at large-scale simulations relevant to engineering applications, with high dimensional input parameters.

Sensitivity analyses and other forward UQ problems (e.g., estimation of failure probabilities, rare events,. . .) depends on the input uncertainty model. Most often, for convenience or because of the lack of data, the independence of the uncertain inputs is assumed. In the Team, we are investigating approaches dedicated to a) the construction of uncertainty models that integrate the available information and expert knowledge(s) in a consistent and objective fashion. To this end, several mathematical frameworks are already available, e.g the maximum entropy principle, likelihood maximization and moment matching methods, but their application to real engineering problems remains scarce and their systematic use raises multiple challenges, both to construct the uncertainty model and to solve the related UQ problems (forward and backward). Because of the importance of the available data and expertise to build the model, the contributions of the Team in these areas depend on the needs and demands of end-users and industrial partners.

To mitigate computational complexity, the Team is exploring multi-fidelity approaches in the context of expensive simulations. We combine predictions of models with different levels of discretizations and physical simplifications to construct, at a controlled cost, reliable surrogate models of simulation outputs or directly objective functions and possibly constraints, to enable the resolution of robust optimization and stochastic inverse problems. Again, one difficulty to be addressed by the Team is the design of the computer experiments to obtain the best multi-fidelity model at the lowest cost (of for a prescribed computational budgets), with respect to the end use of the model. The last point is particularly challenging as it calls for accuracy for output values that are usually unknown a priori but must be estimated as the model construction proceeds.

DISCO Project-Team

3. Research Program

3.1. Analysis of interconnected systems

The major questions considered are those of the characterization of the stability (also including the problems of sensitivity compared to the variations of the parameters) and the determination of stabilizing controllers of interconnected dynamic systems. In many situations, the dynamics of the interconnections can be naturally modelled by systems with delays (constant, distributed or time-varying delays) possibly of fractional order. In other cases, partial differential equations (PDE) models can be better represented or approximated by using systems with delays. Our expertise on this subject, on both time and frequency domain methods, allows us to challenge difficult problems (e.g. systems with an infinite number of unstable poles).

- Robust stability of linear systems

  Within an interconnection context, lots of phenomena are modelled directly or after an approximation by delay systems. These systems may have constant delays, time-varying delays, distributed delays ...

  For various infinite-dimensional systems, particularly delay and fractional systems, input-output and time-domain methods are jointly developed in the team to characterize stability. This research is developed at four levels: analytic approaches (H_∞-stability, BIBO-stablity, robust stability, robustness metrics) [1], [2], [6], [7], symbolic computation approaches (SOS methods are used for determining easy-to-check conditions which guarantee that the poles of a given linear system are not in the closed right half-plane, certified CAD techniques), numerical approaches (root-loci, continuation methods) and by means of softwares developed in the team [6], [7].

- Robustness/fragility of biological systems

  Deterministic biological models describing, for instance, species interactions, are frequently composed of equations with important disturbances and poorly known parameters. To evaluate the impact of the uncertainties, we use the techniques of designing of global strict Lyapunov functions or functional developed in the team.

  However, for other biological systems, the notion of robustness may be different and this question is still in its infancy (see, e.g. [57]). Unlike engineering problems where a major issue is to maintain stability in the presence of disturbances, a main issue here is to maintain the system response in the presence of disturbances. For instance, a biological network is required to keep its functioning in case of a failure of one of the nodes in the network. The team, which has a strong expertise in robustness for engineering problems, aims at contributing at the develpment of new robustness metrics in this biological context.

3.2. Stabilization of interconnected systems

- Linear systems: Analytic and algebraic approaches are considered for infinite-dimensional linear systems studied within the input-output framework.

  In the recent years, the Youla-Kučera parametrization (which gives the set of all stabilizing controllers of a system in terms of its coprime factorizations) has been the cornerstone of the success of the H_∞-control since this parametrization allows one to rewrite the problem of finding the optimal stabilizing controllers for a certain norm such as H_∞ or H_2 as affine, and thus, convex problem.

  A central issue studied in the team is the computation of such factorizations for a given infinite-dimensional linear system as well as establishing the links between stabilizability of a system for a certain norm and the existence of coprime factorizations for this system. These questions are fundamental for robust stabilization problems [1], [2].

We also consider simultaneous stabilization since it plays an important role in the study of reliable stabilization, i.e. in the design of controllers which stabilize a finite family of plants describing a system during normal operating conditions and various failed modes (e.g. loss of sensors or actuators, changes in operating points). Moreover, we investigate strongly stabilizable systems, namely systems which can be stabilized by stable controllers, since they have a good ability to track reference inputs and, in practice, engineers are reluctant to use unstable controllers especially when the system is stable.

- Nonlinear systems

  In any physical systems a feedback control law has to account for limitation stemming from safety, physical or technological constraints. Therefore, any realistic control system analysis and design has to account for these limitations appearing mainly from sensors and actuators nonlinearities and from the regions of safe operation in the state space. This motivates the study of linear systems with more realistic, thus complex, models of actuators. These constraints appear as nonlinearities as saturation and quantization in the inputs of the system [10].

  The project aims at developing robust stabilization theory and methods for important classes of nonlinear systems that ensure good controller performance under uncertainty and time delays. The main techniques include techniques called backstepping and forwarding, contructions of strict Lyapunov functions through so-called "strictification" approaches [4] and construction of Lyapunov-Krasovskii functionals [5], [6], [7] or or Lyapunov functionals for PDE systems [9].

3.3. Synthesis of reduced complexity controllers

- PID controllers

  Even though the synthesis of control laws of a given complexity is not a new problem, it is still open, even for finite-dimensional linear systems. Our purpose is to search for good families of “simple” (e.g. low order) controllers for infinite-dimensional dynamical systems. Within our approach, PID candidates are first considered in the team [2], [60].

  For interconnected systems appearing in teleoperation applications, such as the steer-by-wire, Proportional-Derivative laws are simple control strategies allowing to reproduce the efforts in both ends of the teleoperation system. However, due to delays introduced in the communication channels these strategies may result in loss of closed loop stability or in performance degradation when compared to the system with a mechanical link (no communication channel). In this context we search for non-linear proportional and derivative gains to improve performance. This is assessed in terms of reduction of overshoot and guaranteed convergence rates.

- Delayed feedback

  Control systems often operate in the presence of delays, primarily due to the time it takes to acquire the information needed for decision-making, to create control decisions and to execute these decisions. Commonly, such a time delay induces desynchronizing and/or destabilizing effects on the dynamics. However, some recent studies have emphasized that the delay may have a stabilizing effect in the control design. In particular, the closed-loop stability may be guaranteed precisely by the existence of the delay. The interest of considering such control laws lies in the simplicity of the controller as well as in its easy practical implementation. It is intended by the team members to provide a unified approach for the design of such stabilizing control laws for finite and infinite dimensional plants [3], [8].

- Finite Time and Interval Observers for nonlinear systems

  We aim to develop techniques of construction of output feedbacks relying on the design of observers. The objectives pertain to the design of robust control laws which converge in finite time, the construction of intervals observers which ensure that the solutions belong to guaranteed intervals, continuous/discrete observers for systems with discrete measurements and observers for systems with switches.

Finally, the development of algorithms based on both symbolic computation and numerical methods, and their implementations in dedicated Scilab/Matlab/Maple toolboxes are important issues in the project.

ECUADOR Project-Team

3. Research Program

3.1. Algorithmic Differentiation

Participants: Laurent Hascoët, Valérie Pascual.

algorithmic differentiation (AD, aka Automatic Differentiation) Transformation of a program, that returns a new program that computes derivatives of the initial program, i.e. some combination of the partial derivatives of the program's outputs with respect to its inputs.

adjoint Mathematical manipulation of the Partial Differential Equations that define a problem, obtaining new differential equations that define the gradient of the original problem's solution.

checkpointing General trade-off technique, used in adjoint AD, that trades duplicate execution of a part of the program to save some memory space that was used to save intermediate results.

Algorithmic Differentiation (AD) differentiates *programs*. The input of AD is a source program P that, given some $X \in \mathbb{R}^n$, returns some $Y = F(X) \in \mathbb{R}^m$, for a differentiable F. AD generates a new source program P' that, given X, computes some derivatives of F [4].

Any execution of P amounts to a sequence of instructions, which is identified with a composition of vector functions. Thus, if

$$\begin{array}{lll} P & \text{runs} & \{I_1; I_2; \cdots I_p;\}, \\ F & \text{then is} & f_p \circ f_{p-1} \circ \cdots \circ f_1, \end{array} \tag{5}$$

where each f_k is the elementary function implemented by instruction I_k. AD applies the chain rule to obtain derivatives of F. Calling X_k the values of all variables after instruction I_k, i.e. $X_0 = X$ and $X_k = f_k(X_{k-1})$, the Jacobian of F is

$$F'(X) = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \tag{6}$$

which can be mechanically written as a sequence of instructions I'_k. This can be generalized to higher level derivatives, Taylor series, etc. Combining the I'_k with the control of P yields P', and therefore this differentiation is piecewise.

The above computation of $F'(X)$, albeit simple and mechanical, can be prohibitively expensive on large codes. In practice, many applications only need cheaper projections of $F'(X)$ such as:

- **Sensitivities**, defined for a given direction $\dot{X}$ in the input space as:

$$F'(X).\dot{X} = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \,.\, \dot{X} \quad . \tag{7}$$

  This expression is easily computed from right to left, interleaved with the original program instructions. This is the *tangent mode* of AD.

- **Adjoints**, defined after transposition (F'^*), for a given weighting $\overline{Y}$ of the outputs as:

$$F'^*(X).\overline{Y} = f'^*_1(X_0).f'^*_2(X_1). \cdots .f'^*_{p-1}(X_{p-2}).f'^*_p(X_{p-1}).\overline{Y} \quad . \tag{8}$$

This expression is most efficiently computed from right to left, because matrix×vector products are cheaper than matrix×matrix products. This is the *adjoint mode* of AD, most effective for optimization, data assimilation [31], adjoint problems [25], or inverse problems.

Adjoint AD builds a very efficient program [27], which computes the gradient in a time independent from the number of parameters n. In contrast, computing the same gradient with the *tangent mode* would require running the tangent differentiated program n times.

However, the X_k are required in the *inverse* of their computation order. If the original program *overwrites* a part of X_k, the differentiated program must restore X_k before it is used by $f'^*_{k+1}(X_k)$. Therefore, the central research problem of adjoint AD is to make the X_k available in reverse order at the cheapest cost, using strategies that combine storage, repeated forward computation from available previous values, or even inverted computation from available later values.

Another research issue is to make the AD model cope with the constant evolution of modern language constructs. From the old days of Fortran77, novelties include pointers and dynamic allocation, modularity, structured data types, objects, vectorial notation and parallel programming. We keep developing our models and tools to handle these new constructs.

3.2. Static Analysis and Transformation of programs

Participants: Laurent Hascoët, Valérie Pascual.

abstract syntax tree Tree representation of a computer program, that keeps only the semantically significant information and abstracts away syntactic sugar such as indentation, parentheses, or separators.

control flow graph Representation of a procedure body as a directed graph, whose nodes, known as basic blocks, each contain a sequence of instructions and whose arrows represent all possible control jumps that can occur at run-time.

abstract interpretation Model that describes program static analysis as a special sort of execution, in which all branches of control switches are taken concurrently, and where computed values are replaced by abstract values from a given *semantic domain*. Each particular analysis gives birth to a specific semantic domain.

data flow analysis Program analysis that studies how a given property of variables evolves with execution of the program. Data Flow analysis is static, therefore studying all possible run-time behaviors and making conservative approximations. A typical data-flow analysis is to detect, at any location in the source program, whether a variable is initialized or not.

The most obvious example of a program transformation tool is certainly a compiler. Other examples are program translators, that go from one language or formalism to another, or optimizers, that transform a program to make it run better. AD is just one such transformation. These tools share the technological basis that lets them implement the sophisticated analyses [14] required. In particular there are common mathematical models to specify these analyses and analyze their properties.

An important principle is *abstraction*: the core of a compiler should not bother about syntactic details of the compiled program. The optimization and code generation phases must be independent from the particular input programming language. This is generally achieved using language-specific *front-ends*, language-independent *middle-ends*, and target-specific *back-ends*. In the middle-end, analysis can concentrate on the semantics of a reduced set of constructs. This analysis operates on an abstract representation of programs made of one *call graph*, whose nodes are themselves *flow graphs* whose nodes (*basic blocks*) contain abstract *syntax trees* for the individual atomic instructions. To each level are attached symbol tables, nested to capture scoping.

Static program analysis can be defined on this internal representation, which is largely language independent. The simplest analyses on trees can be specified with inference rules [18], [28], [15]. But many *data-flow analyses* are more complex, and better defined on graphs than on trees. Since both call graphs and flow graphs may be cyclic, these global analyses will be solved iteratively. *Abstract Interpretation* [19] is a theoretical framework to study complexity and termination of these analyses.

Data flow analyses must be carefully designed to avoid or control combinatorial explosion. At the call graph level, they can run bottom-up or top-down, and they yield more accurate results when they take into account the different call sites of each procedure, which is called *context sensitivity*. At the flow graph level, they can run forwards or backwards, and yield more accurate results when they take into account only the possible execution flows resulting from possible control, which is called *flow sensitivity*.

Even then, data flow analyses are limited, because they are static and thus have very little knowledge of actual run-time values. Far before reaching the very theoretical limit of *undecidability*, one reaches practical limitations to how much information one can infer from programs that use arrays [34], [20] or pointers. Therefore, conservative *over-approximations* must be made, leading to derivative code less efficient than ideal.

3.3. Algorithmic Differentiation and Scientific Computing

Participants: Alain Dervieux, Laurent Hascoët, Bruno Koobus, Eléonore Gauci, Emmanuelle Itam, Olivier Allain, Stephen Wornom.

linearization In Scientific Computing, the mathematical model often consists of Partial Differential Equations, that are discretized and then solved by a computer program. Linearization of these equations, or alternatively linearization of the computer program, predict the behavior of the model when small perturbations are applied. This is useful when the perturbations are effectively small, as in acoustics, or when one wants the sensitivity of the system with respect to one parameter, as in optimization.

adjoint state Consider a system of Partial Differential Equations that define some characteristics of a system with respect to some parameters. Consider one particular scalar characteristic. Its sensitivity (or gradient) with respect to the parameters can be defined by means of *adjoint* equations, deduced from the original equations through linearization and transposition. The solution of the adjoint equations is known as the adjoint state.

Scientific Computing provides reliable simulations of complex systems. For example it is possible to *simulate* the steady or unsteady 3D air flow around a plane that captures the physical phenomena of shocks and turbulence. Next comes *optimization*, one degree higher in complexity because it repeatedly simulates and applies gradient-based optimization steps until an optimum is reached. The next sophistication is *robustness*, that detects undesirable solutions which, although maybe optimal, are very sensitive to uncertainty on design parameters or on manufacturing tolerances. This makes second derivatives come into play. Similarly *Uncertainty Quantification* can use second derivatives to evaluate how uncertainty on the simulation inputs imply uncertainty on its outputs.

To obtain this gradient and possibly higher derivatives, we advocate adjoint AD (*cf*3.1) of the program that discretizes and solves the direct system. This gives the exact gradient of the discrete function computed by the program, which is quicker and more sound than differentiating the original mathematical equations [25]. Theoretical results [24] guarantee convergence of these derivatives when the direct program converges. This approach is highly mechanizable. However, it requires careful study and special developments of the AD model [29], [32] to master possibly heavy memory usage. Among these additional developments, we promote in particular specialized AD models for Fixed-Point iterations [26], [17], efficient adjoints for linear algebra operators such as solvers, or exploitation of parallel properties of the adjoint code.

ELAN Project-Team

3. Research Program

3.1. Discrete modeling of slender elastic structures

For the last 15 years, we have investigated new discrete models for solving the Kirchhoff dynamic equations for thin elastic rods [10], [12], [15]. All our models share a curvature-based spatial discretization, allowing them to capture inextensibility of the rod intrinsically, without the need for adding any kinematic constraint. Moreover, elastic forces boil down to linear terms in the dynamic equations, making them well-suited for implicit integration. Interestingly, our discretization methodology can be interpreted from two different points-of-views. From the finite-elements point-of-view, our strain-based discrete schemes can be seen as discontinuous Galerkin methods of zero and first orders. From the multibody system dynamics point of view, our discrete models can be interpreted as deformable Lagrangian systems in finite dimension, for which a dedicated community has started to grow recently [37]. We note that adopting the multibody system dynamics point of view helped us formulate a linear-time integration scheme [11], which had only be investigated in the case of multibody rigid bodies dynamics so far.

3.1.1. High-order spatial discretization schemes for rods, ribbons and shells

Our goal is to investigate similar high-order modeling strategies for surfaces, in particular for the case of inextensible ribbons and shells. Elastic ribbons have been scarcely studied in the past, but they are nowadays drawing more and more the attention from physicists [25], [34]. Their numerical modeling remains an open challenge. In contrast to ribbons, a huge litterature exists for shells, both from a theoretical and numerical viewpoints (see, e.g., [29], [16]). However, no real consensus has been obtained so far about a unified nonlinear shell theory able to support large displacements. In [13] we have started building an inextensible shell patch by taking as degrees of freedom the curvatures of its mid-surface, expressed in the local frame. As in the super-helix model, we show that when taking curvatures uniform over the element, each term of the equations of motion may be computed in closed-form; besides, the geometry of the element corresponds to a cylinder patch at each time step. Compared to the 1D (rod) case however, some difficulties arise in the 2D (plate/shell) case, where compatibility conditions are to be treated carefully.

3.1.2. Numerical continuation of rod equilibria in the presence of unilateral constraints

In Alejandro Blumentals' PhD thesis [14], we have adopted an optimal control point of view on the static problem of thin elastic rods, and we have shown that direct discretization methods [0] are particularly well-suited for dealing with scenarios involving both bilateral and unilateral constraints (such as contact). We would like to investigate how our formulations extend to continuation problems, where the goal is to follow a certain branch of equilibria when the rod is subject to some varying constraints (such as one fixed end being applied a constant rotation). To the best of our knowledge, classical continuation methods used for rods [26] are not able to deal with non-persistent or sliding contact.

3.2. Discrete and continuous modeling of frictional contact

Most popular approaches in Computer Graphics and Mechanical Engineering consist in assuming that the objects in contact are locally compliant, allowing them to slightly penetrate each other. This is the principle of penalty-based methods (or molecular dynamics), which consists in adding mutual repulsive forces of the form $k\,f(\delta)$, where δ is the penetration depth detected at current time step [17], [33]. Though simple to implement and computationally efficient, the penalty-based method often fails to prevent excessive penetration of the contacting objects, which may prove fatal in the case of thin objects as those may just end up traversing each other. One solution might be to set the stiffness factor k to a large enough value, however this causes the introduction of parasitical high frequencies and calls for very small integration steps [9]. Penalty-based approaches are thus generally not satisfying for ensuring robust contact handling.

[0] Within this optimal control framework, our previous curvature-based methods can actually be interpreted as a special case of direct single shooting methods.

In the same vein, the friction law between solid objects, or within a yield-stress fluid (used to model foam, sand, or cement, which, unlike water, cannot flow beyond a certain threshold), is commonly modeled using a regularized friction law (sometimes even with simple viscous forces), for the sake of simplicity and numerical tractability (see e.g., [36], [28]). Such a model cannot capture the threshold effect that characterizes friction between contacting solids or within a yield-stress fluid. The nonsmooth transition between sticking and sliding is however responsible for significant visual features, such as the complex patterns resting on the outer surface of hair, the stable formation of sand piles, or typical stick-slip instabilities occurring during motion.

The search for a realistic, robust and stable frictional contact method encouraged us to depart from those, and instead to focus on rigid contact models coupled to the exact nonsmooth Coulomb law for friction (and respectively, to the exact nonsmooth Drucker-Prager law in the case of a fluid), which better integrate the effects of frictional contact at the macroscopic scale. This motivation was the sense of the hiring of F. Bertails-Descoubes in 2007 in the Inria/LJK BIPOP team, specialized in nonsmooth mechanics and related convex optimization methods. In the line of F. Bertails-Descoubes's work performed in the BIPOP team, the ELAN team keeps on including some active research on the finding of robust frictional contact algorithms specialized for slender deformable structures.

3.2.1. Optimized algorithms for large nodal systems in frictional contact

In the fiber assembly case, the resulting mass matrix M is block-diagonal, so that the Delassus operator can be computed in an efficient way by leveraging sparse-block computations [18]. This justifies solving the reduced discrete frictional contact problem where primary unknowns are forces, as usually advocated in nonsmooth mechanics [31]. For cloth however, where primal variables (nodal velocities of the cloth mesh) are all interconnected via elasticity through implicit forces, the method developed above is computationally inefficient. Indeed, the matrix M (only block-sparse, but not block-diagonal) is costly to invert for large systems and its inverse is dense. Recently, we have leveraged the fact that generalized velocities of the system are 3D velocities, which simplifies the discrete contact problem when contacts occur at the nodes. Combined with a multiresolution strategy, we have devised an algorithm able to capture exact Coulomb friction constraints at contact, while retaining computational efficiency [32]. This work also supports cloth self-contact and cloth multilayering. How to enrich the interaction model with, e.g., cohesion, remains an open question. The experimental validation of our frictional contact model is also one of our goals in the medium run.

3.2.2. Continuum modeling of granular and fibrous media

Though we have recently made progress on the continuum formulation and solving of granular materials in Gilles Daviet's PhD thesis [22], [20], [19], we are still far from a continuum description of a macroscopic dry fibrous medium such as hair. One key ingredient that we have not been considering in our previous models is the influence of air inside divided materials. Typically, air plays a considerable role in hair motion. To advance in that direction, we have started to look at a diphasic fluid representation of granular matter, where a Newtonian fluid and the solid phase are fully coupled, while the nonsmooth Drucker-Prager rheology for the solid phase is enforced implicitly [21]. This first approach could be a starting point for modeling immersed granulars in a liquid, or ash clouds, for instance.

A long path then remains to be achieved, if one wants to take into account long fibers instead of isotropic grains in the solid phase. How to couple the fiber elasticity with our current formulation remains a challenging problem.

3.3. Inverse design of slender elastic structures [ERC Gem]

With the considerable advance of automatic image-based capture in Computer Vision and Computer Graphics these latest years, it becomes now affordable to acquire quickly and precisely the full 3D geometry of many mechanical objects featuring intricate shapes. Yet, while more and more geometrical data get collected and shared among the communities, there is currently very little study about how to infer the underlying mechanical properties of the captured objects merely from their geometrical configurations.

An important challenge consists in developing a non-invasive method for inferring the mechanical properties of complex objects from a minimal set of geometrical poses, in order to predict their dynamics. In contrast to classical inverse reconstruction methods, our claim is that 1/ the mere geometrical shape of physical objects reveals a lot about their underlying mechanical properties and 2/ this property can be fully leveraged for a wide range of objects featuring rich geometrical configurations, such as slender structures subject to contact and friction (e.g., folded cloth or twined filaments).

In addition to significant advances in fast image-based measurement of diverse mechanical materials stemming from physics, biology, or manufacturing, this research is expected in the long run to ease considerably the design of physically realistic virtual worlds, as well as to boost the creation of dynamic human doubles.

To achieve this goal, we shall develop an original inverse modeling strategy based upon the following research topics:

3.3.1. Design of well-suited discrete models for slender structures

We believe that the quality of the upstream, reference physics-based model is essential to the effective connection between geometry and mechanics. Typically, such a model should properly account for the nonlinearities due to large displacements of the structures, as well as to the nonsmooth effects typical of contact and friction.

It should also be parameterized and discretized in such a way that inversion gets simplified mathematically, possibly avoiding the huge cost of large and nonconvex optimization. In that sense, unlike concurrent methods which impose inverse methods to be compatible with a generic physics-based model, we instead advocate the design of specific physics-based models which are tailored for the inversion process.

More precisely, from our experience on fiber modeling, we believe that reduced Lagrangian models, based on a minimal set of coordinates and physical parameters (as opposed to maximal coordinates models such as mass-springs), are particularly well-suited for inversion and physical interpretation of geometrical data [24], [23]. Furthermore, choosing a high-order coordinate system (e.g., curvatures instead of angles) allows for a precise handling of curved boundaries and contact geometry, as well as the simplification of constitutive laws (which are transformed into a linear equation in the case of rods). We are currently investigating high-order discretization schemes for elastic ribbons and developable shells [13].

3.3.2. Static inversion of physical objects from geometrical poses

We believe that pure static inversion may by itself reveal many insights regarding a range of parameters such as the undeformed configuration of the object, some material parameters or contact forces.

The typical settings that we consider is composed of, on the one hand, a reference mechanical model of the object of interest, and on the other hand a single or a series of complete geometrical poses corresponding each to a static equilibrium. The core challenge consists in analyzing theoretically and practically the amount of information that can be gained from one or several geometrical poses, and to understand how the fundamental under-determinacy of the inverse problem can be reduced, for each unknown quantity (parameter or force) at play. Both the equilibrium condition and the stability criterion of the equilibrium are leveraged towards this goal. On the theoretical side, we have recently shown that a given 3D curve always matches the centerline of an isotropic suspended Kirchhoff rod at equilibrium under gravity, and that the natural configuration of the rod is unique once material parameters (mass, Young modulus) are fixed [1]. On the practical side, we have recently devised a robust algorithm to find a valid natural configuration for a discrete shell to match a given surface under gravity and frictional contact forces [3]. Unlike rods however, shells can have multiple inverse (natural) configurations. Choosing among the multiple solutions based on some selection criteria is an open challenge. Another open issue, in all cases, is the theoretical characterization of material parameters allowing the equilibrium to be stable.

3.3.3. Dynamic inversion of physical objects from geometrical poses

To refine the solution subspaces searched for in the static case and estimate dynamic parameters (e.g., some damping coefficients), a dynamic inversion process accounting for the motion of the object of interest is necessary.

In contrast to the static case where we can afford to rely on exact geometrical poses, our analysis in the dynamic case will have to take into account the imperfect quality of input data with possible missing parts or outliers. One interesting challenge will be to combine our high-order discretized physics-based model together with the acquisition process in order to refine both the parameter estimation and the geometrical acquisition.

3.3.4. Experimental validation with respect to real data

The goal will be to confront the theories developed above to real experiments. Compared to the statics, the dynamic case will be particularly involving as it will be highly dependent on the quality of input data as well as the accuracy of the motion predicted by our physics-based simulators. Such experiments will not only serve to refine our direct and inverse models, but will also be leveraged to improve the 3D geometrical acquisition of moving objects. Besides, once validation will be performed, we shall work on the setting up of new non-invasive measurement protocols to acquire physical parameters of slender structures from a minimal amount of geometrical configurations.

FACTAS Project-Team

3. Research Program

3.1. Introduction

Within the extensive field of inverse problems, much of the research by Factas deals with reconstructing solutions of classical elliptic PDEs from their boundary behavior. Perhaps the simplest example lies with harmonic identification of a stable linear dynamical system: the transfer-function f can be evaluated at a point $i\omega$ of the imaginary axis from the response to a periodic input at frequency ω. Since f is holomorphic in the right half-plane, it satisfies there the Cauchy-Riemann equation $\overline{\partial} f = 0$, and recovering f amounts to solve a Dirichlet problem which can be done in principle using, *e.g.* the Cauchy formula.

Practice is not nearly as simple, for f is only measured pointwise in the pass-band of the system which makes the problem ill-posed [70]. Moreover, the transfer function is usually sought in specific form, displaying the necessary physical parameters for control and design. For instance if f is rational of degree n, then $\overline{\partial} f = \sum_1^n a_j \delta_{z_j}$ where the z_j are its poles and δ_{z_j} is a Dirac unit mass at z_j. Thus, to find the domain of holomorphy (*i.e.* to locate the z_j) amounts to solve a (degenerate) free-boundary inverse problem, this time on the left half-plane. To address such questions, the team has developed a two-step approach as follows.

> **Step 1:** To determine a complete model, that is, one which is defined at every frequency, in a sufficiently versatile function class (*e.g.* Hardy spaces). This ill-posed issue requires regularization, for instance constraints on the behavior at non-measured frequencies.
>
> **Step 2:** To compute a reduced order model. This typically consists of rational approximation of the complete model obtained in step 1, or phase-shift thereof to account for delays. We emphasize that deriving a complete model in step 1 is crucial to achieve stability of the reduced model in step 2.

Step 1 relates to extremal problems and analytic operator theory, see Section 3.3.1 . Step 2 involves optimization, and some Schur analysis to parametrize transfer matrices of given Mc-Millan degree when dealing with systems having several inputs and outputs, see Section 3.3.2.2 . It also makes contact with the topology of rational functions, in particular to count critical points and to derive bounds, see Section 3.3.2 . Step 2 raises further issues in approximation theory regarding the rate of convergence and the extent to which singularities of the approximant (*i.e.* its poles) tend to singularities of the approximated function; this is where logarithmic potential theory becomes instrumental, see Section 3.3.3 .

Applying a realization procedure to the result of step 2 yields an identification procedure from incomplete frequency data which was first demonstrated in [76] to tune resonant microwave filters. Harmonic identification of nonlinear systems around a stable equilibrium can also be envisaged by combining the previous steps with exact linearization techniques from [34].

A similar path can be taken to approach design problems in the frequency domain, replacing the measured behavior by some desired behavior. However, describing achievable responses in terms of the design parameters is often cumbersome, and most constructive techniques rely on specific criteria adapted to the physics of the problem. This is especially true of filters, the design of which traditionally appeals to polynomial extremal problems [72], [57]. To this area, Apics contributed the use of Zolotarev-like problems for multi-band synthesis, although we presently favor interpolation techniques in which parameters arise in a more transparent manner, as well as convex relaxation of hyperbolic approximation problems, see Sections 3.2.2 and 6.2.2 .

The previous example of harmonic identification quickly suggests a generalization of itself. Indeed, on identifying $\mathbb{C}$ with $\mathbb{R}^2$, holomorphic functions become conjugate-gradients of harmonic functions, so that harmonic identification is, after all, a special case of a classical issue: to recover a harmonic function on a domain from partial knowledge of the Dirichlet-Neumann data; when the portion of boundary where data are not available is itself unknown, we meet a free boundary problem. This framework for 2-D non-destructive control was first advocated in [62] and subsequently received considerable attention. It makes clear how

to state similar problems in higher dimensions and for more general operators than the Laplacian, provided solutions are essentially determined by the trace of their gradient on part of the boundary which is the case for elliptic equations [0] [32], [80]. Such questions are particular instances of the so-called inverse potential problem, where a measure μ has to be recovered from the knowledge of the gradient of its potential (*i.e.*, the field) on part of a hypersurface (a curve in 2-D) encompassing the support of μ. For Laplace's operator, potentials are logarithmic in 2-D and Newtonian in higher dimensions. For elliptic operators with non constant coefficients, the potential depends on the form of fundamental solutions and is less manageable because it is no longer of convolution type. Nevertheless it is a useful concept bringing perspective on how problems could be raised and solved, using tools from harmonic analysis.

Inverse potential problems are severely indeterminate because infinitely many measures within an open set of $\mathbb{R}^n$ produce the same field outside this set; this phenomenon is called *balayage* [69]. In the two steps approach previously described, we implicitly removed this indeterminacy by requiring in step 1 that the measure be supported on the boundary (because we seek a function holomorphic throughout the right half-space), and by requiring in step 2 that the measure be discrete in the left half-plane (in fact: a finite sum of point masses $\sum_1^N a_j \delta_{z_j}$). The discreteness assumption also prevails in 3-D inverse source problems, see Section 4.3 . Conditions that ensure uniqueness of the solution to the inverse potential problem are part of the so-called regularizing assumptions which are needed in each case to derive efficient algorithms.

To recap, the gist of our approach is to approximate boundary data by (boundary traces of) fields arising from potentials of measures with specific support. This differs from standard approaches to inverse problems, where descent algorithms are applied to integration schemes of the direct problem; in such methods, it is the equation which gets approximated (in fact: discretized).

Along these lines, Factas advocates the use of steps 1 and 2 above, along with some singularity analysis, to approach issues of nondestructive control in 2-D and 3-D [2], [41], [45]. The team is currently engaged in the generalization to inverse source problems for the Laplace equation in 3-D, to be described further in Section 3.2.1 . There, holomorphic functions are replaced by harmonic gradients; applications are to inverse source problems in neurosciences (in particular in EEG/MEG) and inverse magnetization problems in geosciences, see Section 4.3 .

The approximation-theoretic tools developed by Apics and now by Factas to handle issues mentioned so far are outlined in Section 3.3 . In Section 3.2 to come, we describe in more detail which problems are considered and which applications are targeted.

Note that the Inria project-team Apics reached the end of its life cycle by the end of 2017. The proposal for our new team Factas was processed by the CEP (Comité des Équipes-Projets) of the Research Center in 2018, and approved by the head of the Institute in 2019.

3.2. Range of inverse problems

3.2.1. Elliptic partial differential equations (PDE)

Participants: Paul Asensio, Laurent Baratchart, Sylvain Chevillard, Juliette Leblond, Masimba Nemaire, Konstantinos Mavreas.

By standard properties of conjugate differentials, reconstructing Dirichlet-Neumann boundary conditions for a function harmonic in a plane domain, when these conditions are already known on a subset E of the boundary, is equivalent to recover a holomorphic function in the domain from its boundary values on E. This is the problem raised on the half-plane in step 1 of Section 3.1 . It makes good sense in holomorphic Hardy spaces where functions are entirely determined by their values on boundary subsets of positive linear measure, which is the framework for Problem (P) that we set up in Section 3.3.1 . Such issues naturally

[0] There is a subtle difference here between dimension 2 and higher. Indeed, a function holomorphic on a plane domain is defined by its non-tangential limit on a boundary subset of positive linear measure, but there are non-constant harmonic functions in the 3-D ball, C^1 up to the boundary sphere, yet having vanishing gradient on a subset of positive measure of the sphere. Such a "bad" subset, however, cannot have interior points on the sphere.

arise in nondestructive testing of 2-D (or 3-D cylindrical) materials from partial electrical measurements on the boundary. For instance, the ratio between the tangential and the normal currents (the so-called Robin coefficient) tells one about corrosion of the material. Thus, solving Problem (P) where ψ is chosen to be the response of some uncorroded piece with identical shape yields non destructive testing of a potentially corroded piece of material, part of which is inaccessible to measurements. This was an initial application of holomorphic extremal problems to non-destructive control [55], [58].

Another application by the team deals with non-constant conductivity over a doubly connected domain, the set E being now the outer boundary. Measuring Dirichlet-Neumann data on E, one wants to recover level lines of the solution to a conductivity equation, which is a so-called free boundary inverse problem. For this, given a closed curve inside the domain, we first quantify how constant the solution on this curve. To this effect, we state and solve an analog of Problem (P), where the constraint bears on the real part of the function on the curve (it should be close to a constant there), in a Hardy space of a conjugate Beltrami equation, of which the considered conductivity equation is the compatibility condition (just like the Laplace equation is the compatibility condition of the Cauchy-Riemann system). Subsequently, a descent algorithm on the curve leads one to improve the initial guess. For example, when the domain is regarded as separating the edge of a tokamak's vessel from the plasma (rotational symmetry makes this a 2-D situation), this method can be used to estimate the shape of a plasma subject to magnetic confinement.

This was actually carried out in collaboration with CEA (French nuclear agency) and the University of Nice (JAD Lab.), to data from *Tore Supra* in [61]. The procedure is fast because no numerical integration of the underlying PDE is needed, as an explicit basis of solutions to the conjugate Beltrami equation in terms of Bessel functions was found in this case. Generalizing this approach in a more systematic manner to free boundary problems of Bernoulli type, using descent algorithms based on shape-gradient for such approximation-theoretic criteria, is an interesting prospect to the team.

The piece of work we just mentioned requires defining and studying Hardy spaces of conjugate Beltrami equations, which is an interesting topic. For Sobolev-smooth coefficients of exponent greater than 2, they were investigated in [6], [35]. The case of the critical exponent 2 is treated in [31], which apparently provides the first example of well-posed Dirichlet problem in the non-strictly elliptic case: the conductivity may be unbounded or zero on sets of zero capacity and, accordingly, solutions need not be locally bounded. More importantly perhaps, the exponent 2 is also the key to a corresponding theory on very general (still rectifiable) domains in the plane, as coefficients of pseudo-holomorphic functions obtained by conformal transformation onto a disk are merely of L^2-class in general, even if the initial problem deals with coefficients of L^r-class for some $r > 2$. Such generalizations are now under study within the team.

Generalized Hardy classes as above are used in [32] where we address the uniqueness issue in the classical Robin inverse problem on a Lipschitz domain of $\Omega \subset \mathbb{R}^n$, $n \geq 2$, with uniformly bounded Robin coefficient, L^2 Neumann data and conductivity of Sobolev class $W^{1,r}(\Omega)$, $r > n$. We show that uniqueness of the Robin coefficient on a subset of the boundary, given Cauchy data on the complementary part, does hold in dimension $n = 2$, thanks to a unique continuation result, but needs not hold in higher dimension. In higher dimension, this raises an open issue on harmonic gradients, namely whether the positivity of the Robin coefficient is compatible with identical vanishing of the boundary gradient on a subset of positive measure.

The 3-D version of step 1 in Section 3.1 is another subject investigated by Factas: to recover a harmonic function (up to an additive constant) in a ball or a half-space from partial knowledge of its gradient. This prototypical inverse problem (*i.e.* inverse to the Cauchy problem for the Laplace equation) often recurs in electromagnetism. At present, Factas is involved with solving instances of this inverse problem arising in two fields, namely medical imaging *e.g.* for electroencephalography (EEG) or magneto-encephalography (MEG), and paleomagnetism (recovery of rocks magnetization) [2], [37], see Section 6.1 . In this connection, we collaborate with two groups of partners: Athena Inria project-team and INS (Institut de Neurosciences des Systèmes, http://ins.univ-amu.fr/), hospital la Timone, Aix-Marseille Univ., on the one hand, Geosciences Lab. at MIT and Cerege CNRS Lab. on the other hand. The question is considerably more difficult than its 2-D counterpart, due mainly to the lack of multiplicative structure for harmonic gradients. Still, substantial progress has been made over the last years using methods of harmonic analysis and operator theory.

The team is further concerned with 3-D generalizations and applications to non-destructive control of step 2 in Section 3.1 . A typical problem is here to localize inhomogeneities or defaults such as cracks, sources or occlusions in a planar or 3-dimensional object, knowing thermal, electrical, or magnetic measurements on the boundary. These defaults can be expressed as a lack of harmonicity of the solution to the associated Dirichlet-Neumann problem, thereby posing an inverse potential problem in order to recover them. In 2-D, finding an optimal discretization of the potential in Sobolev norm amounts to solve a best rational approximation problem, and the question arises as to how the location of the singularities of the approximant (*i.e.* its poles) reflects the location of the singularities of the potential (*i.e.* the defaults we seek). This is a fairly deep issue in approximation theory, to which Apics contributed convergence results for certain classes of fields expressed as Cauchy integrals over extremal contours for the logarithmic potential [8], [38], [52]. Initial schemes to locate cracks or sources *via* rational approximation on planar domains were obtained this way [41], [45], [55]. It is remarkable that finite inverse source problems in 3-D balls, or more general algebraic surfaces, can be approached using these 2-D techniques upon slicing the domain into planar sections [9], [42]. More precisely, each section cuts out a planar domain, the boundary of which carries data which can be proved to match an algebraic function. The singularities of this algebraic function are not located at the 3-D sources, but are related to them: the section contains a source if and only if some function of the singularities in that section meets a relative extremum. Using bisection it is thus possible to determine an extremal place along all sections parallel to a given plane direction, up to some threshold which has to be chosen small enough that one does not miss a source. This way, we reduce the original source problem in 3-D to a sequence of inverse poles and branchpoints problems in 2-D. This bottom line generates a steady research activity within Factas, and again applications are sought to medical imaging and geosciences, see Sections 4.3 , 4.2 and 6.1 .

Conjectures may be raised on the behavior of optimal potential discretization in 3-D, but answering them is an ambitious program still in its infancy.

3.2.2. Systems, transfer and scattering

Participants: Laurent Baratchart, Sylvain Chevillard, Adam Cooman, Martine Olivi, Fabien Seyfert.

Through contacts with CNES (French space agency), members of the team became involved in identification and tuning of microwave electromagnetic filters used in space telecommunications, see Section 4.4 . The initial problem was to recover, from band-limited frequency measurements, physical parameters of the device under examination. The latter consists of interconnected dual-mode resonant cavities with negligible loss, hence its scattering matrix is modeled by a 2×2 unitary-valued matrix function on the frequency line, say the imaginary axis to fix ideas. In the bandwidth around the resonant frequency, a modal approximation of the Helmholtz equation in the cavities shows that this matrix is approximately rational, of Mc-Millan degree twice the number of cavities.

This is where system theory comes into play, through the so-called *realization* process mapping a rational transfer function in the frequency domain to a state-space representation of the underlying system of linear differential equations in the time domain. Specifically, realizing the scattering matrix allows one to construct a virtual electrical network, equivalent to the filter, the parameters of which mediate in between the frequency response and the geometric characteristics of the cavities (*i.e.* the tuning parameters).

Hardy spaces provide a framework to transform this ill-posed issue into a series of regularized analytic and meromorphic approximation problems. More precisely, the procedure sketched in Section 3.1 goes as follows:

1. infer from the pointwise boundary data in the bandwidth a stable transfer function (*i.e.* one which is holomorphic in the right half-plane), that may be infinite dimensional (numerically: of high degree). This is done by solving a problem analogous to (P) in Section 3.3.1 , while taking into account prior knowledge on the decay of the response outside the bandwidth, see [13] for details.
2. A stable rational approximation of appropriate degree to the model obtained in the previous step is performed. For this, a descent method on the compact manifold of inner matrices of given size and degree is used, based on an original parametrization of stable transfer functions developed within the team [27], [13].

3. Realizations of this rational approximant are computed. To be useful, they must satisfy certain constraints imposed by the geometry of the device. These constraints typically come from the coupling topology of the equivalent electrical network used to model the filter. This network is composed of resonators, coupled according to some specific graph. This realization step can be recast, under appropriate compatibility conditions [56], as solving a zero-dimensional multivariate polynomial system. To tackle this problem in practice, we use Gröbner basis techniques and continuation methods which team up in the Dedale-HF software (see Section 3.4.2).

We recently started a collaboration with the Chinese Hong Kong University on the topic of frequency depending couplings appearing in the equivalent circuits we compute [19] continuing our work [1] on wide-band design applications.

Factas also investigates issues pertaining to design rather than identification. Given the topology of the filter, a basic problem in this connection is to find the optimal response subject to specifications that bear on rejection, transmission and group delay of the scattering parameters. Generalizing the classical approach based on Chebyshev polynomials for single band filters, we recast the problem of multi-band response synthesis as a generalization of the classical Zolotarev min-max problem for rational functions [26] [12]. Thanks to quasi-convexity, the latter can be solved efficiently using iterative methods relying on linear programming. These were implemented in the software easy-FF (see easy-FF). Currently, the team is engaged in the synthesis of more complex microwave devices like multiplexers and routers, which connect several filters through wave guides. Schur analysis plays an important role here, because scattering matrices of passive systems are of Schur type (*i.e.* contractive in the stability region). The theory originates with the work of I. Schur [75], who devised a recursive test to check for contractivity of a holomorphic function in the disk. The so-called Schur parameters of a function may be viewed as Taylor coefficients for the hyperbolic metric of the disk, and the fact that Schur functions are contractions for that metric lies at the root of Schur's test. Generalizations thereof turn out to be efficient to parametrize solutions to contractive interpolation problems [28]. Dwelling on this, Factas contributed differential parametrizations (atlases of charts) of lossless matrix functions [27], [71], [66] which are fundamental to our rational approximation software RARL2 (see Section 3.4.5). Schur analysis is also instrumental to approach de-embedding issues, and provides one with considerable insight into the so-called matching problem. The latter consists in maximizing the power a multiport can pass to a given load, and for reasons of efficiency it is all-pervasive in microwave and electric network design, *e.g.* of antennas, multiplexers, wifi cards and more. It can be viewed as a rational approximation problem in the hyperbolic metric, and the team presently deals with this hot topic using contractive interpolation with constraints on boundary peak points, within the framework of the (defense funded) ANR Cocoram, see Sections 6.2 .

In recent years, our attention was driven by CNES and UPV (Bilbao) to questions about stability of high-frequency amplifiers. Contrary to previously discussed devices, these are *active* components. The response of an amplifier can be linearized around a set of primary current and voltages, and then admittances of the corresponding electrical network can be computed at various frequencies, using the so-called harmonic balance method. The initial goal is to check for stability of the linearized model, so as to ascertain existence of a well-defined working state. The network is composed of lumped electrical elements namely inductors, capacitors, negative *and* positive resistors, transmission lines, and controlled current sources. Our research so far has focused on describing the algebraic structure of admittance functions, so as to set up a function-theoretic framework where the two-steps approach outlined in Section 3.1 can be put to work. The main discovery is that the unstable part of each partial transfer function is rational and can be computed by analytic projection, see Section 6.3 . We now start investigating the linearized harmonic transfer-function around a periodic cycle, to check for stability under non necessarily small inputs. This topic generates the doctoral work of S. Fueyo.

3.3. Approximation

Participants: Laurent Baratchart, Sylvain Chevillard, Juliette Leblond, Martine Olivi, Fabien Seyfert.

3.3.1. Best analytic approximation

In dimension 2, the prototypical problem to be solved in step 1 of Section 3.1 may be described as: given a domain $D \subset \mathbb{R}^2$, to recover a holomorphic function from its values on a subset K of the boundary of D. For the discussion it is convenient to normalize D, which can be done by conformal mapping. So, in the simply connected case, we fix D to be the unit disk with boundary unit circle T. We denote by H^p the Hardy space of exponent p, which is the closure of polynomials in $L^p(T)$-norm if $1 \leq p < \infty$ and the space of bounded holomorphic functions in D if $p = \infty$. Functions in H^p have well-defined boundary values in $L^p(T)$, which makes it possible to speak of (traces of) analytic functions on the boundary.

To find an analytic function g in D matching some measured values f approximately on a sub-arc K of T, we formulate a constrained best approximation problem as follows.

> (P) Let $1 \leq p \leq \infty$, K a sub-arc of T, $f \in L^p(K)$, $\psi \in L^p(T \setminus K)$ and $M > 0$; find a function $g \in H^p$ such that $\|g - \psi\|_{L^p(T \setminus K)} \leq M$ and $g - f$ is of minimal norm in $L^p(K)$ under this constraint.

Here ψ is a reference behavior capturing *a priori* assumptions on the behavior of the model off K, while M is some admissible deviation thereof. The value of p reflects the type of stability which is sought and how much one wants to smooth out the data. The choice of L^p classes is suited to handle pointwise measurements.

To fix terminology, we refer to (P) as a *bounded extremal problem*. As shown in [40], [43], [49], the solution to this convex infinite-dimensional optimization problem can be obtained when $p \neq 1$ upon iterating with respect to a Lagrange parameter the solution to spectral equations for appropriate Hankel and Toeplitz operators. These spectral equations involve the solution to the special case $K = T$ of (P), which is a standard extremal problem [64]:

> (P_0) Let $1 \leq p \leq \infty$ and $\varphi \in L^p(T)$; find a function $g \in H^p$ such that $g - \varphi$ is of minimal norm in $L^p(T)$.

In the case $p = 1$, partial results are known but computational issues remain open.

Various modifications of (P) can be tailored to meet specific needs. For instance when dealing with lossless transfer functions (see Section 4.4), one may want to express the constraint on $T \setminus K$ in a pointwise manner: $|g - \psi| \leq M$ a.e. on $T \setminus K$, see [44]. In this form, the problem comes close to (but still is different from) H^∞ frequency optimization used in control [67], [74]. One can also impose bounds on the real or imaginary part of $g - \psi$ on $T \setminus K$, which is useful when considering Dirichlet-Neumann problems.

The analog of Problem (P) on an annulus, K being now the outer boundary, can be seen as a means to regularize a classical inverse problem occurring in nondestructive control, namely to recover a harmonic function on the inner boundary from Dirichlet-Neumann data on the outer boundary (see Sections 3.2.1 , 4.3 , 6.1.3). It may serve as a tool to approach Bernoulli type problems, where we are given data on the outer boundary and we *seek the inner boundary*, knowing it is a level curve of the solution. In this case, the Lagrange parameter indicates how to deform the inner contour in order to improve data fitting. Similar topics are discussed in Section 3.2.1 for more general equations than the Laplacian, namely isotropic conductivity equations of the form $\mathrm{div}(\sigma \nabla u) = 0$ where σ is no longer constant (*i.e.*, varies in the space). Then, the Hardy spaces in Problem (P) are those of a so-called conjugate Beltrami equation: $\overline{\partial} f = \nu \overline{\partial f}$ [68], which are studied for $1 < p < \infty$ in [6], [31], [35] and [59]. Expansions of solutions needed to constructively handle such issues in the specific case of linear fractional conductivities (occurring for instance in plasma shaping) have been expounded in [61].

Though originally considered in dimension 2, Problem (P) carries over naturally to higher dimensions where analytic functions get replaced by gradients of harmonic functions. Namely, given some open set $\Omega \subset \mathbb{R}^n$ and some $\mathbb{R}^n$-valued vector field V on an open subset O of the boundary of Ω, we seek a harmonic function in Ω whose gradient is close to V on O.

When Ω is a ball or a half-space, a substitute for holomorphic Hardy spaces is provided by the Stein-Weiss Hardy spaces of harmonic gradients [78]. Conformal maps are no longer available when $n > 2$, so that Ω can no longer be normalized. More general geometries than spheres and half-spaces have not been much studied so far.

On the ball, the analog of Problem (P) is

> (P_1) Let $1 \leq p \leq \infty$ and $B \subset \mathbb{R}^n$ the unit ball. Fix O an open subset of the unit sphere $S \subset \mathbb{R}^n$. Let further $V \in L^p(O)$ and $W \in L^p(S \setminus O)$ be $\mathbb{R}^n$-valued vector fields. Given $M > 0$, find a harmonic gradient $G \in H^p(B)$ such that $\|G - W\|_{L^p(S \setminus O)} \leq M$ and $G - V$ is of minimal norm in $L^p(O)$ under this constraint.

When $p = 2$, Problem (P_1) was solved in [2] as well as its analog on a shell, when the tangent component of V is a gradient (when O is Lipschitz the general case follows easily from this). The solution extends the work in [40] to the 3-D case, using a generalization of Toeplitz operators. The case of the shell was motivated by applications to the processing of EEG data. An important ingredient is a refinement of the Hodge decomposition, that we call the *Hardy-Hodge* decomposition, allowing us to express a $\mathbb{R}^n$-valued vector field in $L^p(S)$, $1 < p < \infty$, as the sum of a vector field in $H^p(B)$, a vector field in $H^p(\mathbb{R}^n \setminus \overline{B})$, and a tangential divergence free vector field on S; the space of such divergence-free fields is denoted by $D(S)$. If $p = 1$ or $p = \infty$, L^p must be replaced by the real Hardy space or the space of functions with bounded mean oscillation. More generally this decomposition, which is valid on any sufficiently smooth surface (see Section 6.1), seems to play a fundamental role in inverse potential problems. In fact, it was first introduced formally on the plane to describe silent magnetizations supported in $\mathbb{R}^2$ (*i.e.* those generating no field in the upper half space) [37].

Just like solving problem (P) appeals to the solution of problem (P_0), our ability to solve problem (P_1) will depend on the possibility to tackle the special case where $O = S$:

> (P_2) Let $1 \leq p \leq \infty$ and $V \in L^p(S)$ be a $\mathbb{R}^n$-valued vector field. Find a harmonic gradient $G \in H^p(B)$ such that $\|G - V\|_{L^p(S)}$ is minimum.

Problem (P_2) is simple when $p = 2$ by virtue of the Hardy-Hodge decomposition together with orthogonality of $H^2(B)$ and $H^2(\mathbb{R}^n \setminus \overline{B})$, which is the reason why we were able to solve (P_1) in this case. Other values of p cannot be treated as easily and are still under investigation, especially the case $p = \infty$ which is of particular interest and presents itself as a 3-D analog to the Nehari problem [73].

Companion to problem (P_2) is problem (P_3) below.

> (P_3) Let $1 \leq p \leq \infty$ and $V \in L^p(S)$ be a $\mathbb{R}^n$-valued vector field. Find $G \in H^p(B)$ and $D \in D(S)$ such that $\|G + D - V\|_{L^p(S)}$ is minimum.

Note that (P_2) and (P_3) are identical in 2-D, since no non-constant tangential divergence-free vector field exists on T. It is no longer so in higher dimension, where both (P_2) and (P_3) arise in connection with inverse potential problems in divergence form, like source recovery in electro/magneto encephalography and paleomagnetism, see Sections 3.2.1 and 4.3 .

3.3.2. Best meromorphic and rational approximation

The techniques set forth in this section are used to solve step 2 in Section 3.2 and they are instrumental to approach inverse boundary value problems for the Poisson equation $\Delta u = \mu$, where μ is some (unknown) measure.

3.3.2.1. Scalar meromorphic and rational approximation

We put R_N for the set of rational functions with at most N poles in D. By definition, meromorphic functions in $L^p(T)$ are (traces of) functions in $H^p + R_N$.

A natural generalization of problem (P_0) is:

> (P_N) Let $1 \leq p \leq \infty$, $N \geq 0$ an integer, and $f \in L^p(T)$; find a function $g_N \in H^p + R_N$ such that $g_N - f$ is of minimal norm in $L^p(T)$.

Only for $p = \infty$ and f continuous is it known how to solve (P_N) in semi-closed form. The unique solution is given by AAK theory (named after Adamjan, Arov and Krein), which connects the spectral decomposition of Hankel operators with best approximation [73].

The case where $p = 2$ is of special importance for it reduces to rational approximation. Indeed, if we write the Hardy decomposition $f = f^+ + f^-$ where $f^+ \in H^2$ and $f^- \in H^2(\mathbb{C} \setminus \overline{D})$, then $g_N = f^+ + r_N$ where r_N is a best approximant to f^- from R_N in $L^2(T)$. Moreover, r_N has no pole outside D, hence it is a *stable* rational approximant to f^-. However, in contrast to the case where $p = \infty$, this best approximant may *not* be unique.

The Miaou project (predecessor of Apics) already designed a dedicated steepest-descent algorithm for the case $p = 2$ whose convergence to a *local minimum* is guaranteed; the algorithm ha evolved over years and still now, it seems to be the only procedure meeting this property. This gradient algorithm proceeds recursively with respect to N on a compactification of the parameter space [33]. Although it has proved to be effective in all applications carried out so far (see Sections 4.3 , 4.4), it is still unknown whether the absolute minimum can always be obtained by choosing initial conditions corresponding to *critical points* of lower degree (as is done by the RARL2 software, Section 3.4.5).

In order to establish global convergence results, Apics has undertaken a deeper study of the number and nature of critical points (local minima, saddle points, ...), in which tools from differential topology and operator theory team up with classical interpolation theory [46], [48]. Based on this work, uniqueness or asymptotic uniqueness of the approximant was proved for certain classes of functions like transfer functions of relaxation systems (*i.e.* Markov functions) [50] and more generally Cauchy integrals over hyperbolic geodesic arcs [53]. These are the only results of this kind. Research by Apics on this topic remained dormant for a while by reasons of opportunity, but revisiting the work [29] in higher dimension is a worthy and timely endeavor today. Meanwhile, an analog to AAK theory was carried out for $2 \leq p < \infty$ in [49]. Although not as effective computationally, it was recently used to derive lower bounds [5]. When $1 \leq p < 2$, problem (P_N) is still quite open.

A common feature to the above-mentioned problems is that critical point equations yield non-Hermitian orthogonality relations for the denominator of the approximant. This stresses connections with interpolation, which is a standard way to build approximants, and in many respects best or near-best rational approximation may be regarded as a clever manner to pick interpolation points. This was exploited in [54], [51], and is used in an essential manner to assess the behavior of poles of best approximants to functions with branched singularities, which is of particular interest for inverse source problems (*cf.* Sections 3.4.3 and 6.1).

In higher dimensions, the analog of Problem (P_N) is best approximation of a vector field by gradients of discrete potentials generated by N point masses. This basic issue is by no means fully understood, and it is an exciting field of research. It is connected with certain generalizations of Toeplitz or Hankel operators, and with constructive approaches to so-called weak factorizations for real Hardy functions [60].

Besides, certain constrained rational approximation problems, of special interest in identification and design of passive systems, arise when putting additional requirements on the approximant, for instance that it should be smaller than 1 in modulus (*i.e.* a Schur function). In particular, Schur interpolation lately received renewed attention from the team, in connection with matching problems. There, interpolation data are subject to a well-known compatibility condition (positive definiteness of the so-called Pick matrix), and the main difficulty is to put interpolation points on the boundary of D while controlling both the degree and the extremal points (peak points for the modulus) of the interpolant. Results obtained by Apics in this direction generalize a variant of contractive interpolation with degree constraint as studied in [65]. We mention that contractive interpolation with nodes approaching the boundary has been a subsidiary research topic by the team in the past, which plays an interesting role in the spectral representation of certain non-stationary stochastic processes [36], [39].

3.3.2.2. Matrix-valued rational approximation

Matrix-valued approximation is necessary to handle systems with several inputs and outputs but it generates additional difficulties as compared to scalar-valued approximation, both theoretically and algorithmically. In the matrix case, the McMillan degree (*i.e.* the degree of a minimal realization in the System-Theoretic sense) generalizes the usual notion of degree for rational functions. For instance when poles are simple, the McMillan degree is the sum of the ranks of the residues.

The basic problem that we consider now goes as follows: *let* $\mathcal{F} \in (H^2)^{m\times l}$ *and* n *an integer; find a rational matrix of size* $m \times l$ *without poles in the unit disk and of McMillan degree at most* n *which is nearest possible to* $\mathcal{F}$ *in* $(H^2)^{m\times l}$. Here the L^2 norm of a matrix is the square root of the sum of the squares of the norms of its entries.

The scalar approximation algorithm derived in [33] and mentioned in Section 3.3.2.1 generalizes to the matrix-valued situation [63]. The first difficulty here is to parametrize inner matrices (*i.e.* matrix-valued functions analytic in the unit disk and unitary on the unit circle) of given McMillan degree degree n. Indeed, inner matrices play the role of denominators in fractional representations of transfer matrices (using the so-called Douglas-Shapiro-Shields factorization). The set of inner matrices of given degree is a smooth manifold that allows one to use differential tools as in the scalar case. In practice, one has to produce an atlas of charts (local parametrizations) and to handle changes of charts in the course of the algorithm. Such parametrization can be obtained using interpolation theory and Schur-type algorithms, the parameters of which are vectors or matrices ([27], [66], [71]). Some of these parametrizations are also interesting to compute realizations and achieve filter synthesis ([66], [71]). The rational approximation software "RARL2" developed by the team is described in Section 3.4.5 .

Difficulties relative to multiple local minima of course arise in the matrix-valued case as well, and deriving criteria that guarantee uniqueness is even more difficult than in the scalar case. The case of rational functions of degree n or small perturbations thereof (the consistency problem) was solved in [47]. Matrix-valued Markov functions are the only known example beyond this one [30].

Let us stress that RARL2 seems the only algorithm handling rational approximation in the matrix case that demonstrably converges to a local minimum while meeting stability constraints on the approximant. It is still a working pin of many developments by Factas on frequency optimization and design.

3.3.3. Behavior of poles of meromorphic approximants

Participant: Laurent Baratchart.

We refer here to the behavior of poles of best meromorphic approximants, in the L^p-sense on a closed curve, to functions f defined as Cauchy integrals of complex measures whose support lies inside the curve. Normalizing the contour to be the unit circle T, we are back to Problem (P_N) in Section 3.3.2.1 ; invariance of the latter under conformal mapping was established in [45]. Research so far has focused on functions whose singular set inside the contour is polar, meaning that the function can be continued analytically (possibly in a multiple-valued manner) except over a set of logarithmic capacity zero.

Generally speaking in approximation theory, assessing the behavior of poles of rational approximants is essential to obtain error rates as the degree goes large, and to tackle constructive issues like uniqueness. However, as explained in Section 3.2.1 , the original twist by Apics, now Factas, is to consider this issue also as a means to extract information on singularities of the solution to a Dirichlet-Neumann problem. The general theme is thus: *how do the singularities of the approximant reflect those of the approximated function?* This approach to inverse problem for the 2-D Laplacian turns out to be attractive when singularities are zero- or one-dimensional (see Section 4.3). It can be used as a computationally cheap initial condition for more precise but much heavier numerical optimizations which often do not even converge unless properly initialized. As regards crack detection or source recovery, this approach boils down to analyzing the behavior of best meromorphic approximants of given pole cardinality to a function with branch points, which is the prototype of a polar singular set. For piecewise analytic cracks, or in the case of sources, we were able to prove ([8], [45], [38]), that the poles of the approximants accumulate, when the degree goes large, to some extremal cut of minimum weighted logarithmic capacity connecting the singular points of the crack, or the sources [41]. Moreover, the asymptotic density of the poles turns out to be the Green equilibrium distribution on this cut in D, therefore it charges the singular points if one is able to approximate in sufficiently high degree (this is where the method could fail, because high-order approximation requires rather precise data).

The case of two-dimensional singularities is still an outstanding open problem.

It is remarkable that inverse source problems inside a sphere or an ellipsoid in 3-D can be approached with such 2-D techniques, as applied to planar sections, see Section 6.1 . The technique is implemented in the software FindSources3D, see Section 3.4.3 .

3.4. Software tools of the team

In addition to the above-mentioned research activities, Factas develops and maintains a number of long-term software tools that either implement and illustrate effectiveness of the algorithms theoretically developed by the team or serve as tools to help further research by team members. We present briefly the most important of them.

3.4.1. Pisa

KEYWORDS: Electrical circuit - Stability

FUNCTIONAL DESCRIPTION: To minimise prototyping costs, the design of analog circuits is performed using computer-aided design tools which simulate the circuit's response as accurately as possible.

Some commonly used simulation tools do not impose stability, which can result in costly errors when the prototype turns out to be unstable. A thorough stability analysis is therefore a very important step in circuit design. This is where pisa is used.

pisa is a Matlab toolbox that allows designers of analog electronic circuits to determine the stability of their circuits in the simulator. It analyses the impedance presented by a circuit to determine the circuit's stability. When an instability is detected, pisa can estimate location of the unstable poles to help designers fix their stability issue.

RELEASE FUNCTIONAL DESCRIPTION: First version

- Authors: Adam Cooman, David Martinez Martinez, Fabien Seyfert and Martine Olivi
- Contact: Fabien Seyfert
- Publications: Model-Free Closed-Loop Stability Analysis: A Linear Functional Approach - On Transfer Functions Realizable with Active Electronic Components
- URL: https://project.inria.fr/pisa

3.4.2. DEDALE-HF

SCIENTIFIC DESCRIPTION

Dedale-HF consists in two parts: a database of coupling topologies as well as a dedicated predictor-corrector code. Roughly speaking each reference file of the database contains, for a given coupling topology, the complete solution to the coupling matrix synthesis problem (C.M. problem for short) associated to particular filtering characteristics. The latter is then used as a starting point for a predictor-corrector integration method that computes the solution to the C.M. corresponding to the user-specified filter characteristics. The reference files are computed off-line using Gröbner basis techniques or numerical techniques based on the exploration of a monodromy group. The use of such continuation techniques, combined with an efficient implementation of the integrator, drastically reduces the computational time.

Dedale-HF has been licensed to, and is currently used by TAS-Espana

FUNCTIONAL DESCRIPTION

Dedale-HF is a software dedicated to solve exhaustively the coupling matrix synthesis problem in reasonable time for the filtering community. Given a coupling topology, the coupling matrix synthesis problem consists in finding all possible electromagnetic coupling values between resonators that yield a realization of given filter characteristics. Solving the latter is crucial during the design step of a filter in order to derive its physical dimensions, as well as during the tuning process where coupling values need to be extracted from frequency measurements.

- Participant: Fabien Seyfert
- Contact: Fabien Seyfert
- URL: http://www-sop.inria.fr/apics/Dedale/

3.4.3. FindSources3D

KEYWORDS: Health - Neuroimaging - Visualization - Compilers - Medical - Image - Processing
FindSources3D is a software program dedicated to the resolution of inverse source problems in electroencephalography (EEG). From pointwise measurements of the electrical potential taken by electrodes on the scalp, FindSources3D estimates pointwise dipolar current sources within the brain in a spherical model.

After a first data transmission "cortical mapping" step, it makes use of best rational approximation on 2-D planar cross-sections and of the software RARL2 in order to locate singularities. From those planar singularities, the 3-D sources are estimated in a last step, see [9].

The present version of FindSources3D (called FindSources3D-bolis) provides a modular, ergonomic, accessible and interactive platform, with a convenient graphical interface for EEG medical imaging. Modularity is now granted (using the tools dtk, Qt, with compiled Matlab libraries). It offers a detailed and nice visualization of data and tuning parameters, processing steps, and of the computed results (using VTK).

A new version is being developed that will incorporate a first Singular Value Decomposition (SVD) step in order to be able to handle time dependent data and to find the corresponding principal static components.

- Participants: Juliette Leblond, Maureen Clerc (team Athena, Inria Sophia), Jean-Paul Marmorat, Théodore Papadopoulo (team Athena).
- Contact: Juliette Leblond
- URL: http://www-sop.inria.fr/apics/FindSources3D/en/index.html

3.4.4. PRESTO-HF

SCIENTIFIC DESCRIPTION

For the matrix-valued rational approximation step, Presto-HF relies on RARL2. Constrained realizations are computed using the Dedale-HF software. As a toolbox, Presto-HF has a modular structure, which allows one for example to include some building blocks in an already existing software.

The delay compensation algorithm is based on the following assumption: far off the pass-band, one can reasonably expect a good approximation of the rational components of S11 and S22 by the first few terms of their Taylor expansion at infinity, a small degree polynomial in 1/s. Using this idea, a sequence of quadratic convex optimization problems are solved, in order to obtain appropriate compensations. In order to check the previous assumption, one has to measure the filter on a larger band, typically three times the pass band.

This toolbox has been licensed to (and is currently used by) Thales Alenia Space in Toulouse and Madrid, Thales airborne systems and Flextronics (two licenses). Xlim (University of Limoges) is a heavy user of Presto-HF among the academic filtering community and some free license agreements have been granted to the microwave department of the University of Erlangen (Germany) and the Royal Military College (Kingston, Canada).

FUNCTIONAL DESCRIPTION

Presto-HF is a toolbox dedicated to low-pass parameter identification for microwave filters. In order to allow the industrial transfer of our methods, a Matlab-based toolbox has been developed, dedicated to the problem of identification of low-pass microwave filter parameters. It allows one to run the following algorithmic steps, either individually or in a single stroke:

• Determination of delay components caused by the access devices (automatic reference plane adjustment),

• Automatic determination of an analytic completion, bounded in modulus for each channel,

• Rational approximation of fixed McMillan degree,

• Determination of a constrained realization.

- Participants: Fabien Seyfert, Jean-Paul Marmorat and Martine Olivi
- Contact: Fabien Seyfert
- URL: https://project.inria.fr/presto-hf/

3.4.5. RARL2

Réalisation interne et Approximation Rationnelle L2
SCIENTIFIC DESCRIPTION

The method is a steepest-descent algorithm. A parametrization of MIMO systems is used, which ensures that the stability constraint on the approximant is met. The implementation, in Matlab, is based on state-space representations.

RARL2 performs the rational approximation step in the software tools PRESTO-HF and FindSources3D. It is distributed under a particular license, allowing unlimited usage for academic research purposes. It was released to the universities of Delft and Maastricht (the Netherlands), Cork (Ireland), Brussels (Belgium), Macao (China) and BITS-Pilani Hyderabad Campus (India).
FUNCTIONAL DESCRIPTION

RARL2 is a software for rational approximation. It computes a stable rational L2-approximation of specified order to a given L2-stable (L2 on the unit circle, analytic in the complement of the unit disk) matrix-valued function. This can be the transfer function of a multivariable discrete-time stable system. RARL2 takes as input either:

• its internal realization,

• its first N Fourier coefficients,

• discretized (uniformly distributed) values on the circle. In this case, a least-square criterion is used instead of the L2 norm.

It thus performs model reduction in the first or the second case, and leans on frequency data identification in the third. For band-limited frequency data, it could be necessary to infer the behavior of the system outside the bandwidth before performing rational approximation.

An appropriate Möbius transformation allows to use the software for continuous-time systems as well.

- Participants: Jean-Paul Marmorat and Martine Olivi
- Contact: Martine Olivi
- URL: http://www-sop.inria.fr/apics/RARL2/rarl2.html

3.4.6. Sollya

KEYWORDS: Numerical algorithm - Supremum norm - Curve plotting - Remez algorithm - Code generator - Proof synthesis
FUNCTIONAL DESCRIPTION

Sollya is an interactive tool where the developers of mathematical floating-point libraries (libm) can experiment before actually developing code. The environment is safe with respect to floating-point errors, i.e. the user precisely knows when rounding errors or approximation errors happen, and rigorous bounds are always provided for these errors.

Among other features, it offers a fast Remez algorithm for computing polynomial approximations of real functions and also an algorithm for finding good polynomial approximants with floating-point coefficients to any real function. As well, it provides algorithms for the certification of numerical codes, such as Taylor Models, interval arithmetic or certified supremum norms.

It is available as a free software under the CeCILL-C license.

- Participants: Sylvain Chevillard, Christoph Lauter, Mioara Joldes and Nicolas Jourdan
- Partners: CNRS - ENS Lyon - UCBL Lyon 1
- Contact: Sylvain Chevillard
- URL: http://sollya.gforge.inria.fr/

GAMMA Project-Team (section vide)

GEOSTAT Project-Team

3. Research Program

3.1. General methodology

Fully Developed Turbulence (FDT) Turbulence at very high Reynolds numbers; systems in FDT are beyond deterministic chaos, and symmetries are restored in a statistical sense only, and multi-scale correlated structures are landmarks. Generalizing to more random uncorrelated multi-scale structured turbulent fields.

Compact Representation Reduced representation of a complex signal (dimensionality reduction) from which the whole signal can be reconstructed. The reduced representation can correspond to points randomly chosen, such as in Compressive Sensing, or to geometric localization related to statistical information content (framework of reconstructible systems).

Sparse representation The representation of a signal as a linear combination of elements taken in a dictionary (frame or Hilbertian basis), with the aim of finding as less as possible non-zero coefficients for a large class of signals.

Universality class In theoretical physics, the observation of the coincidence of the critical exponents (behaviour near a second order phase transition) in different phenomena and systems is called universality. Universality is explained by the theory of the renormalization group, allowing for the determination of the changes followed by structured fluctuations under rescaling, a physical system is the stage of. The notion is applicable with caution and some differences to generalized out-of-equilibrium or disordered systems. Non-universal exponents (without definite classes) exist in some universal slowing dynamical phenomena like the glass transition and kindred. As a consequence, different macroscopic phenomena displaying multiscale structures (and their acquisition in the form of complex signals) may be grouped into different sets of generalized classes.

Every signal conveys, as a measure experiment, information on the physical system whose signal is an acquisition of. As a consequence, it seems natural that signal analysis or compression should make use of physical modelling of phenomena: the goal is to find new methodologies in signal processing that goes beyond the simple problem of interpretation. Physics of disordered systems, and specifically physics of (spin) glasses is putting forward new algorithmic resolution methods in various domains such as optimization, compressive sensing etc. with significant success notably for NP hard problem heuristics. Similarly, physics of turbulence introduces phenomenological approaches involving multifractality. Energy cascades are indeed closely related to geometrical manifolds defined through random processes. At these structures' scales, information in the process is lost by dissipation (close to the lower bound of inertial range). However, all the cascade is encoded in the geometric manifolds, through long or short distance correlations depending on cases. How do these geometrical manifold structures organize in space and time, in other words, how does the scale entropy cascades itself ? To unify these two notions, a description in term of free energy of a generic physical model is sometimes possible, such as an elastic interface model in a random nonlinear energy landscape : This is for instance the correspondence between compressible stochastic Burgers equation and directed polymers in a disordered medium. Thus, trying to unlock the fingerprints of cascade-like structures in acquired natural signals becomes a fundamental problem, from both theoretical and applicative viewpoints.

To illustrate the general methodology undertaken, let us focus on an example conducted in the study of physiological time series: the analysis of signals recorded from the electrical activity of the heart in the general setting of Atrial Fibrillation (AF). AF is a cardiac arrhythmia characterized by rapid and irregular atrial electrical activity with a high clinical impact on stroke incidence. Best available therapeutic strategies combine pharmacological and surgical means. But when successful, they do not always prevent long-term relapses.

Initial success becomes all the more tricky to achieve as the arrhythmia maintains itself and the pathology evolves into sustained or chronic AF. This raises the open crucial issue of deciphering the mechanisms that govern the onset of AF as well as its perpetuation. We have developed a wavelet-based multi-scale strategy to analyze the electrical activity of human hearts recorded by catheter electrodes, positioned in the coronary sinus (CS), during episodes of chronic AF. We have computed the so-called multifractal spectra using two variants of the wavelet transform modulus maxima method, the moment (partition function) method and the magnitude cumulant method (checking confidence intervals with surrogate data). Application of these methods to long time series recorded in a patient with chronic AF provides quantitative evidence of the multifractal intermittent nature of the electric energy of passing cardiac impulses at low frequencies, *i.e.* for times ($>\sim 0.5$ s) longer than the mean interbeat ($\simeq 10^{-1}$s). We have also reported the results of a two-point magnitude correlation analysis which infers the absence of a multiplicative time-scale structure underlying multifractal scaling. The electric energy dynamics looks like a "multifractal white noise" with quadratic (log-normal) multifractal spectra. *These observations challenge concepts of functional reentrant circuits in mechanistic theories of AF*. A transition is observed in the computed multifractal spectra which group according to two distinct areas, consistently with the anatomical substrate binding to the CS, namely the left atrial posterior wall, and the ligament of Marshall which is innervated by the ANS. These negative results challenge also the existing models, which by principle cannot explain such results. As a consequence, we go beyond the existing models and propose a mathematical model of a denervated heart where the kinetics of gap junction conductance alone induces a desynchronization of the myocardial excitable cells, accounting for the multifractal spectra found experimentally in the left atrial posterior wall area (devoid of ANS influence).

3.2. Turbulence in insterstellar clouds and Earth observation data

The research described in this section is a collaboration effort of GEOSTAT, CNRS LEGOS (Toulouse), CNRS LAM (Marseille Laboratory for Astrophysics), MERCATOR (Toulouse), IIT Roorkee, Moroccan Royal Center for Teledetection (CRST), Moroccan Center for Science CNRST, Rabat University, University of Heidelberg. Researchers involved:

- GEOSTAT: H. Yahia, N. Brodu, K. Daoudi, A. El Aouni, A. Tamim
- CNRS LAB: S. Bontemps, N. Schneider
- CNRS LEGOS: V. Garçon, I. Hernandez-Carrasco, J. Sudre, B. Dewitte
- CNRS LAM: T. Fusco
- CNRST, CRTS, Rabat University: D. Aboutajdine, A. Atillah, K. Minaoui
- Universiy of Heidelberg: C. Garbe

The analysis and modeling of natural phenomena, specially those observed in geophysical sciences and in astronomy, are influenced by statistical and multiscale phenomenological descriptions of turbulence; indeed these descriptions are able to explain the partition of energy within a certain range of scales. A particularly important aspect of the statistical theory of turbulence lies in the discovery that the support of the energy transfer is spatially highly non uniform, in other terms it is *intermittent* [70]. Because of the absence of localization of the Fourier transform, linear methods are not successful to unlock the multiscale structures and cascading properties of variables which are of primary importance as stated by the physics of the phenomena. This is the reason why new approaches, such as DFA (Detrented Fluctuation Analysis), Time-frequency analysis, variations on curvelets [66] etc. have appeared during the last decades. Recent advances in dimensionality reduction, and notably in Compressive Sensing, go beyond the Nyquist rate in sampling theory using nonlinear reconstruction, but data reduction occur at random places, independently of geometric localization of information content, which can be very useful for acquisition purposes, but of lower impact in signal analysis. We are successfully making use of a microcanonical formulation of the multifractal theory, based on predictability and reconstruction, to study the turbulent nature of interstellar molecular or atomic clouds. Another important result obtained in GEOSTAT is the effective use of multiresolution analysis associated to optimal inference along the scales of a complex system. The multiresolution analysis is performed on dimensionless quantities given by the *singularity exponents* which encode properly the geometrical structures

associated to multiscale organization. This is applied successfully in the derivation of high resolution ocean dynamics, or the high resolution mapping of gaseous exchanges between the ocean and the atmosphere; the latter is of primary importance for a quantitative evaluation of global warming. Understanding the dynamics of complex systems is recognized as a new discipline, which makes use of theoretical and methodological foundations coming from nonlinear physics, the study of dynamical systems and many aspects of computer science. One of the challenges is related to the question of *emergence* in complex systems: large-scale effects measurable macroscopically from a system made of huge numbers of interactive agents [31], [61]. Some quantities related to nonlinearity, such as Lyapunov exponents, Kolmogorov-Sinai entropy etc. can be computed at least in the phase space [32]. Consequently, knowledge from acquisitions of complex systems (which include *complex signals*) could be obtained from information about the phase space. A result from F. Takens [67] about strange attractors in turbulence has motivated the theoretical determination of nonlinear characteristics associated to complex acquisitions. Emergence phenomena can also be traced inside complex signals themselves, by trying to localize information content geometrically. Fundamentally, in the nonlinear analysis of complex signals there are broadly two approaches: characterization by attractors (embedding and bifurcation) and time-frequency, multiscale/multiresolution approaches. In real situations, the phase space associated to the acquisition of a complex phenomenon is unknown. It is however possible to relate, inside the signal's domain, local predictability to local reconstruction [13] and to deduce relevant information associated to multiscale geophysical signals [14]. A multiscale organization is a fundamental feature of a complex system, it can be for example related to the cascading properties in turbulent systems. We make use of this kind of description when analyzing turbulent signals: intermittency is observed within the inertial range and is related to the fact that, in the case of FDT (fully developed turbulence), symmetry is restored only in a statistical sense, a fact that has consequences on the quality of any nonlinear signal representation by frames or dictionaries.

The example of FDT as a standard "template" for developing general methods that apply to a vast class of complex systems and signals is of fundamental interest because, in FDT, the existence of a multiscale hierarchy $\mathcal{F}_h$ which is of multifractal nature and geometrically localized can be derived from physical considerations. This geometric hierarchy of sets is responsible for the shape of the computed singularity spectra, which in turn is related to the statistical organization of information content in a signal. It explains scale invariance, a characteristic feature of complex signals. The analogy from statistical physics comes from the fact that singularity exponents are direct generalizations of *critical exponents* which explain the macroscopic properties of a system around critical points, and the quantitative characterization of *universality classes*, which allow the definition of methods and algorithms that apply to general complex signals and systems, and not only turbulent signals: signals which belong to a same universality class share common statistical organization. During the past decades, canonical approaches permitted the development of a well-established analogy taken from thermodynamics in the analysis of complex signals: if $\mathcal{F}$ is the free energy, $\mathcal{T}$ the temperature measured in energy units, $\mathcal{U}$ the internal energy per volume unit $\mathcal{S}$ the entropy and $\widehat{\beta} = 1/\mathcal{T}$, then the scaling exponents associated to moments of intensive variables $p \to \tau_p$ corresponds to $\widehat{\beta}\mathcal{F}$, $\mathcal{U}(\widehat{\beta})$ corresponds to the singularity exponents values, and $\mathcal{S}(\mathcal{U})$ to the singularity spectrum [27]. The research goal is to be able to determine universality classes associated to acquired signals, independently of microscopic properties in the phase space of various complex systems, and beyond the particular case of turbulent data [53].

We show in figure 1 the result of the computation of singularity exponents on an *Herschel* astronomical observation map (the Musca galactic cloud) which has been edge-aware filtered using sparse L^1 filtering to eliminate the cosmic infrared background (or CIB), a type of noise that can modify the singularity spectrum of a signal.

3.3. Causal modeling

The team is working on a new class of models for modeling physical systems, starting from measured data and accounting for their dynamics [40]. The idea is to statistically describe the evolution of a system in terms of causally-equivalent states; states that lead to the same predictions [33]. Transitions between these states can be reconstructed from data, leading to a theoretically-optimal predictive model [63]. In practice, however, no algorithm is currently able to reconstruct these models from data in a reasonable time and without substantial

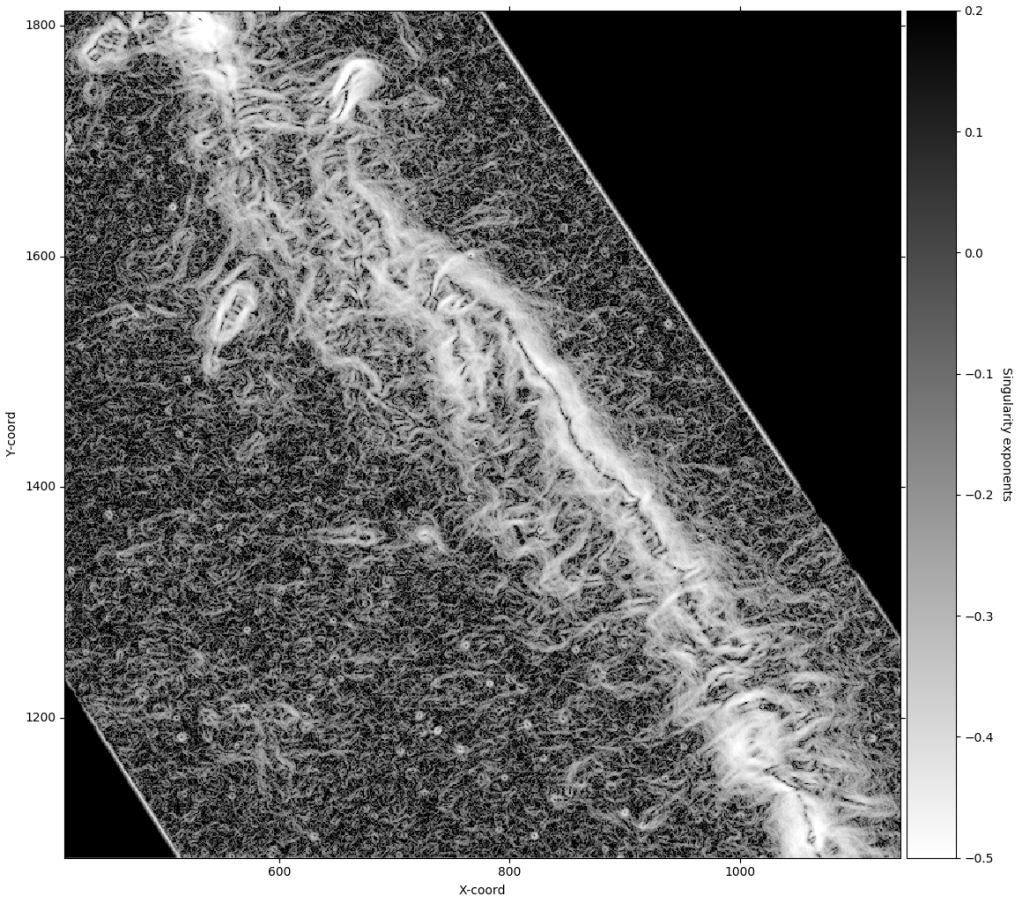


Figure 1. Visualization of the singularity exponents, computed on a edge-aware filtered Musca Herschel observation map .

discrete approximations. Recent progress now allows a continuous formulation of predictive causal models. Within this framework, more efficient algorithms may be found. The broadened class of predictive models promises a new perspective on structural complexity in many applications.

3.4. Speech analysis

Phonetic and sub-phonetic analysis: We developed a novel algorithm for automatic detection of Glottal Closure Instants (GCI) from speech signals using the Microcanonical Multiscale Formalism (MMF). This state of the art algorithm is considered as a reference in this field. We made a Matlab code implementing it available to the community (link). Our approach is based on the Microcanonical Multiscale Formalism. We showed that in the case of clean speech, our algorithm performs almost as well as a recent state-of-the-art method. In presence of different types of noises, we showed that our method is considerably more accurate (particularly for very low SNRs). Moreover, our method has lower computational times does not rely on an estimate of pitch period nor any critical choice of parameters. Using the same MMF, we also developed a method for phonetic segmentation of speech signal. We showed that this method outperforms state of the art ones in term of accuracy and efficiency.

Pathological speech analysis and classification: we made a critical analysis of some widely used methodologies in pathological speech classification. We then introduced some novel methods for extracting some common features used in pathological speech analysis and proposed more robust techniques for classification.

Speech analysis of patients with Parkinsonism: with our collaborators from the Czech Republic, we started preliminary studies of some machine learning issues in the field essentially due the small amount of training data.

3.5. Excitable systems: analysis of physiological time series

The research described in this section is a collaboration effort of GEOSTAT, CNRS LOMA (Laboratoire Ondes et Matière d'Aquitaine) and Laboratory of Physical Foundation of Strength, Institute of Continuous Media Mechanics (Perm, Russia Federation).

AF is an arrhythmia originating in the rapid and irregular electrical activity of the atria (the heart's two upper chambers) that causes their pump function to fail, increasing up to fivefold the risk of embolic stroke. The prevailing electrophysiological concepts describing tachy-arrhythmias are more than a century old. They involve abnormal automaticity and conduction [52]. Initiation and maintenance are thought to arise from a vulnerable substrate prone to the emergence of multiple self-perpetuating reentry circuits, also called "multiple wavelets" [59], [60]. Reentries may be driven structurally, for instance because of locally high fibrous tissue content which badly conducts, or functionally because of high spatial dispersion of decreased refractoriness and APD [58]. The latter is coined the leading circle concept with the clinically more relevant notion of a critical "wavelength" (in fact the length) of the cardiac impulse [26], [65], [62], [30]. The related concept of vulnerability was originally introduced to uncover a physiological substrate evolving from normality to pathology. It was found in vulnerable patients that high rate frequency would invariably lead to functional disorder as cardiac cells would no longer properly adapt their refractoriness [29]. Mathematical models have managed to exhibit likewise phenomena, with the generation of breaking spiral waves in various conditions [49], [54]. The triggering role of abnormal ectopic activity of the pulmonary veins has been demonstrated on patients with paroxysmal AF resistant to drug therapy [48], but its origin still remains poorly understood. This region is highly innervated with sympathetic and parasympathetic stimulation from the ANS [68], [69], [28]. In particular, Coumel et al. [39], [38] have revealed the pathophysiological role of the vagal tone on a vulnerable substrate. It is frequently observed that rapid tachycardia of ectopic origin transits to AF. This is known to result from electrical remodeling. As described for the first time by Allessie et al. [25], remodeling is a transient and reversible process by which the impulse properties such as its refractory period are altered during the course of the arrhythmia, promoting its perpetuation: "AF begets AF" [71]. Under substantial beating rate increase, cells may undergo remodeling to overcome the toxicity of their excessive intercellular calcium loading, by a rapid down regulation (a few minutes) of their L-type calcium membrane current. Moreover, other ionic channel

functions are also modified such as the potassium channel function, inducing a change in the conduction properties including the conduction velocity. The intercellular coupling at the gap junction level shows also alterations of their connexin expression and dispersion.

Wavelet-based methods (WTMM, log-cumulants, two point scale correlations), and confidence statistical methodology, have been applied to catheter recordings in the coronary sinus vein right next to the left atria of a small sample of patients with various conditions, and exhibit clear multifractal scaling without cross-scale correlation, which are coined "multifractal white noise", and that can be grouped according to two anatomical regions. One of our main result was to show that this is incompatible with the common lore for atrial fibrillation based on so-called circuit reentries. We used two declinations of a wavelet-based multiscale method, the moment (partition function) method and the magnitude cumulant method, as originally introduced in the field of fully developed turbulence. In the context of cardiac physiology, this methodology was shown to be valuable in assessing congestive heart failure from the monitoring of sinus heart rate variability [50]. We develop a model such that the substrate function is modulated by the kinetics of conduction. A simple reversible mechanism of short term remodeling under rapid pacing is demonstrated, by which ionic overload acts locally (dynamical feedback) on the kinetics of gap junction conductance. The whole process may propagate and pervade the myocardium via electronic currents, becoming desynchronized. In a new description, we propose that circuit reentries may well exist before the onset of fibrillation, favoring onset but not contributing directly to the onset and perpetuation. By contrast, cell-to-cell coupling is considered fundamentally dynamical. The rationale stems from the observation that multifractal scaling necessitates a high number of degrees of freedom (tending to infinity with system size), which can originate in excitable systems in hyperbolic spatial coupling.

3.6. Data-based identification of characteristic scales and automated modeling

Data are often acquired at the highest possible resolution, but that scale is not necessarily the best for modeling and understanding the system from which data was measured. The intrinsic properties of natural processes do not depend on the arbitrary scale at which data is acquired; yet, usual analysis techniques operate at the acquisition resolution. When several processes interact at different scales, the identification of their characteristic scales from empirical data becomes a necessary condition for properly modeling the system. A classical method for identifying characteristic scales is to look at the work done by the physical processes, the energy they dissipate over time. The assumption is that this work matches the most important action of each process on the studied natural system, which is usually a reasonable assumption. In the framework of time-frequency analysis [45], the power of the signal can be easily computed in each frequency band, itself matching a temporal scale.

However, in open and dissipating systems, energy dissipation is a prerequisite and thus not necessarily the most useful metric to investigate. In fact, most natural, physical and industrial systems we deal with fall in this category, while balanced quasi-static assumptions are practical approximation only for scales well below the characteristic scale of the involved processes. Open and dissipative systems are not locally constrained by the inevitable rise in entropy, thus allowing the maintaining through time of mesoscopic ordered structures. And, according to information theory [47], more order and less entropy means that these structures have a higher information content than the rest of the system, which usually gives them a high functional role.

We propose to identify characteristic scales not only with energy dissipation, as usual in signal processing analysis, but most importantly with information content. Information theory can be extended to look at which scales are most informative (e.g. multi-scale entropy [37], ε-entropy [36]). Complexity measures quantify the presence of structures in the signal (e.g. statistical complexity [42], MPR [56] and others [44]). With these notions, it is already possible to discriminate between random fluctuations and hidden order, such as in chaotic systems [41], [56]. The theory of how information and structures can be defined through scales is not complete yet, but the state of art is promising [43]. Current research in the team focuses on how informative scales can be found using collections of random paths, assumed to capture local structures as they reach out [35].

Building on these notions, it should also possible to fully automate the modeling of a natural system. Once characteristic scales are found, causal relationships can be established empirically. They are then clustered together in internal states of a special kind of Markov models called ϵ-machines [42]. These are known to be the

optimal predictors of a system, with the drawback that it is currently quite complicated to build them properly, except for small system [64]. Recent extensions with advanced clustering techniques [34], [46], coupled with the physics of the studied system (e.g. fluid dynamics), have proved that ϵ-machines are applicable to large systems, such as global wind patterns in the atmosphere [51]. Current research in the team focuses on the use of reproducing kernels, coupled possibly with sparse operators, in order to design better algorithms for ϵ-machines reconstruction. In order to help with this long-term project, a collaboration is ongoing with J. Crutchfield lab at UC Davis.

I4S Project-Team

3. Research Program

3.1. Vibration analysis

In this section, the main features for the key monitoring issues, namely identification, detection, and diagnostics, are provided, and a particular instantiation relevant for vibration monitoring is described.

It should be stressed that the foundations for identification, detection, and diagnostics, are fairly general, if not generic. Handling high order linear dynamical systems, in connection with finite elements models, which call for using subspace-based methods, is specific to vibration-based SHM. Actually, one particular feature of model-based sensor information data processing as exercised in I4S, is the combined use of black-box or semi-physical models together with physical ones. Black-box and semi-physical models are, for example, eigenstructure parameterizations of linear MIMO systems, of interest for modal analysis and vibration-based SHM. Such models are intended to be identifiable. However, due to the large model orders that need to be considered, the issue of model order selection is really a challenge. Traditional advanced techniques from statistics such as the various forms of Akaike criteria (AIC, BIC, MDL, ...) do not work at all. This gives rise to new research activities specific to handling high order models.

Our approach to monitoring assumes that a model of the monitored system is available. This is a reasonable assumption, especially within the SHM areas. The main feature of our monitoring method is its intrinsic ability to the early warning of small deviations of a system with respect to a reference (safe) behavior under usual operating conditions, namely without any artificial excitation or other external action. Such a normal behavior is summarized in a reference parameter vector θ_0, for example a collection of modes and mode-shapes.

3.1.1. Identification

The behavior of the monitored continuous system is assumed to be described by a parametric model $\{\mathbf{P}_\theta\,,\ \theta \in \Theta\}$, where the distribution of the observations $(Z_0, ..., Z_N)$ is characterized by the parameter vector $\theta \in \Theta$.

For reasons closely related to the vibrations monitoring applications, we have been investigating subspace-based methods, for both the identification and the monitoring of the eigenstructure (λ, ϕ_λ) of the state transition matrix F of a linear dynamical state-space system :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & F\ X_k + V_{k+1} \\ Y_k & = & H\ X_k + W_k \end{array} \right. , \tag{9}$$

namely the $(\lambda, \varphi_\lambda)$ defined by :

$$\det\ (F - \lambda\ I) = 0, \quad (F - \lambda\ I)\ \phi_\lambda = 0, \quad \varphi_\lambda \stackrel{\Delta}{=} H\ \phi_\lambda \tag{10}$$

The (canonical) parameter vector in that case is :

$$\theta \stackrel{\Delta}{=} \left(\begin{array}{c} \Lambda \\ \mathrm{vec}\Phi \end{array} \right) \tag{11}$$

where Λ is the vector whose elements are the eigenvalues λ, Φ is the matrix whose columns are the φ_λ's, and vec is the column stacking operator.

Subspace-based methods is the generic name for linear systems identification algorithms based on either time domain measurements or output covariance matrices, in which different subspaces of Gaussian random vectors play a key role [51].

Let $R_i \triangleq \mathbf{E}\left(Y_k\ Y_{k-i}^T\right)$ and:

$$\mathcal{H}_{p+1,q} \triangleq \begin{pmatrix} R_1 & R_2 & \vdots & R_q \\ R_2 & R_3 & \vdots & R_{q+1} \\ \vdots & \vdots & \vdots & \vdots \\ R_{p+1} & R_{p+2} & \vdots & R_{p+q} \end{pmatrix} \triangleq \text{Hank}\left(R_i\right) \tag{12}$$

be the output covariance and Hankel matrices, respectively; and: $G \triangleq \mathbf{E}\left(X_k Y_{k-1}^T\right)$. Direct computations of the R_i's from the equations (4) lead to the well known key factorizations :

$$\begin{aligned} R_i &= HF^{i-1}G \\ \mathcal{H}_{p+1,q} &= \mathcal{O}_{p+1}(H,F)\ \mathcal{C}_q(F,G) \end{aligned} \tag{13}$$

where:

$$\mathcal{O}_{p+1}(H,F) \triangleq \begin{pmatrix} H \\ HF \\ \vdots \\ HF^p \end{pmatrix} \quad \text{and} \quad \mathcal{C}_q(F,G) \triangleq (G\ FG\ \cdots\ F^{q-1}G) \tag{14}$$

are the observability and controllability matrices, respectively. The observation matrix H is then found in the first block-row of the observability matrix $\mathcal{O}$. The state-transition matrix F is obtained from the shift invariance property of $\mathcal{O}$. The eigenstructure (λ, ϕ_λ) then results from (5).

Since the actual model order is generally not known, this procedure is run with increasing model orders.

3.1.2. Detection

Our approach to on-board detection is based on the so-called asymptotic statistical local approach. It is worth noticing that these investigations of ours have been initially motivated by a vibration monitoring application example. It should also be stressed that, as opposite to many monitoring approaches, our method does not require repeated identification for each newly collected data sample.

For achieving the early detection of small deviations with respect to the normal behavior, our approach generates, on the basis of the reference parameter vector θ_0 and a new data record, indicators which automatically perform :

- The early detection of a slight mismatch between the model and the data;
- A preliminary diagnostics and localization of the deviation(s);
- The tradeoff between the magnitude of the detected changes and the uncertainty resulting from the estimation error in the reference model and the measurement noise level.

These indicators are computationally cheap, and thus can be embedded. This is of particular interest in some applications, such as flutter monitoring.

Choosing the eigenvectors of matrix F as a basis for the state space of model (4) yields the following representation of the observability matrix:

$$\mathcal{O}_{p+1}(\theta) = \begin{pmatrix} \Phi \\ \Phi\Delta \\ \vdots \\ \Phi\Delta^p \end{pmatrix} \tag{15}$$

where $\Delta \triangleq \text{diag}(\Lambda)$, and Λ and Φ are as in (6). Whether a nominal parameter θ_0 fits a given output covariance sequence $(R_j)_j$ is characterized by:

$$\mathcal{O}_{p+1}(\theta_0) \text{ and } \mathcal{H}_{p+1,q} \text{ have the same left kernel space.} \tag{16}$$

This property can be checked as follows. From the nominal θ_0, compute $\mathcal{O}_{p+1}(\theta_0)$ using (10), and perform e.g. a singular value decomposition (SVD) of $\mathcal{O}_{p+1}(\theta_0)$ for extracting a matrix U such that:

$$U^T\, U = I_s \text{ and } U^T\, \mathcal{O}_{p+1}(\theta_0) = 0 \tag{17}$$

Matrix U is not unique (two such matrices relate through a post-multiplication with an orthonormal matrix), but can be regarded as a function of θ_0. Then the characterization writes:

$$U(\theta_0)^T\, \mathcal{H}_{p+1,q} = 0 \tag{18}$$

3.1.2.1. Residual associated with subspace identification.

Assume now that *a reference θ_0 and a new sample $Y_1, \cdots, Y_N$ are available.* For checking whether the data agree with θ_0, the idea is to compute the empirical Hankel matrix $\widehat{\mathcal{H}}_{p+1,q}$:

$$\widehat{\mathcal{H}}_{p+1,q} \triangleq \text{Hank}\left(\widehat{R}_i\right), \quad \widehat{R}_i \triangleq 1/(N-i) \sum_{k=i+1}^{N} Y_k\, Y_{k-i}^T \tag{19}$$

and to define the residual vector:

$$\zeta_N(\theta_0) \triangleq \sqrt{N}\ \text{vec}\left(U(\theta_0)^T\, \widehat{\mathcal{H}}_{p+1,q}\right) \tag{20}$$

Let θ be the actual parameter value for the system which generated the new data sample, and $\mathbf{E}_\theta$ be the expectation when the actual system parameter is θ. From (13), we know that $\zeta_N(\theta_0)$ has zero mean when no change occurs in θ, and nonzero mean if a change occurs. Thus $\zeta_N(\theta_0)$ plays the role of a residual.

As in most fault detection approaches, the key issue is to design a *residual*, which is ideally close to zero under normal operation, and has low sensitivity to noises and other nuisance perturbations, but high sensitivity to small deviations, before they develop into events to be avoided (damages, faults, ...). The originality of our approach is to :

- *Design* the residual basically as a *parameter estimating function*,
- *Evaluate* the residual thanks to a kind of central limit theorem, stating that the residual is asymptotically Gaussian and reflects the presence of a deviation in the parameter vector through a change in its own mean vector, which switches from zero in the reference situation to a non-zero value.

The central limit theorem shows [45] that the residual is asymptotically Gaussian :

$$\zeta_N \xrightarrow[N \to \infty]{} \begin{cases} \mathcal{N}(0, \Sigma) & \text{under } \mathbf{P}_{\theta_0} \ , \\ \\ \mathcal{N}(\mathcal{J}\,\eta, \Sigma) & \text{under } \mathbf{P}_{\theta_0 + \eta/\sqrt{N}} \ , \end{cases} \tag{21}$$

where the asymptotic covariance matrix Σ can be estimated, and manifests the deviation in the parameter vector by a change in its own mean value. Then, deciding between $\eta = 0$ and $\eta \neq 0$ amounts to compute the following χ^2-test, provided that $\mathcal{J}$ is full rank and Σ is invertible :

$$\chi^2 = \overline{\zeta}^{\,T}\, \mathbf{F}^{-1}\, \overline{\zeta} \gtrless \lambda \ , \tag{22}$$

where

$$\overline{\zeta} \triangleq \mathcal{J}^T\, \Sigma^{-1}\, \zeta_N \quad \text{and} \quad \mathbf{F} \triangleq \mathcal{J}^T\, \Sigma^{-1}\, \mathcal{J} \ . \tag{23}$$

3.1.3. Diagnostics

A further monitoring step, often called *fault isolation*, consists in determining which (subsets of) components of the parameter vector θ have been affected by the change. Solutions for that are now described. How this relates to diagnostics is addressed afterwards.

The question: *which (subsets of) components of θ have changed ?*, can be addressed using either nuisance parameters elimination methods or a multiple hypotheses testing approach [44].

In most SHM applications, a complex physical system, characterized by a generally non identifiable parameter vector Φ has to be monitored using a simple (black-box) model characterized by an identifiable parameter vector θ. A typical example is the vibration monitoring problem for which complex finite elements models are often available but not identifiable, whereas the small number of existing sensors calls for identifying only simplified input-output (black-box) representations. In such a situation, two different diagnosis problems may arise, namely diagnosis in terms of the black-box parameter θ and diagnosis in terms of the parameter vector Φ of the underlying physical model.

The isolation methods sketched above are possible solutions to the former. Our approach to the latter diagnosis problem is basically a detection approach again, and not a (generally ill-posed) inverse problem estimation approach.

The basic idea is to note that the physical sensitivity matrix writes $\mathcal{J}\,\mathcal{J}_{\Phi\theta}$, where $\mathcal{J}_{\Phi\theta}$ is the Jacobian matrix at Φ_0 of the application $\Phi \mapsto \theta(\Phi)$, and to use the sensitivity test for the components of the parameter vector Φ. Typically this results in the following type of directional test :

$$\chi^2_\Phi = \zeta^T\, \Sigma^{-1}\, \mathcal{J}\, \mathcal{J}_{\Phi\theta}\, \left(\mathcal{J}^T_{\Phi\theta}\, \mathcal{J}^T\, \Sigma^{-1}\, \mathcal{J}\, \mathcal{J}_{\Phi\theta}\right)^{-1} \mathcal{J}^T_{\Phi\theta}\, \mathcal{J}^T\, \Sigma^{-1}\, \zeta \gtrless \lambda \ . \tag{24}$$

It should be clear that the selection of a particular parameterization Φ for the physical model may have a non-negligible influence on such type of tests, according to the numerical conditioning of the Jacobian matrices $\mathcal{J}_{\Phi\theta}$.

3.2. Thermal methods

3.2.1. Infrared thermography and heat transfer

This section introduces the infrared radiation and its link with the temperature, in the next part different measurement methods based on that principle are presented.

3.2.1.1. Infrared radiation

Infrared is an electromagnetic radiation having a wavelength between $0.2\mu m$ and $1\ mm$, this range begin in uv spectrum and it ends on the microwaves domain, see Figure 1 .

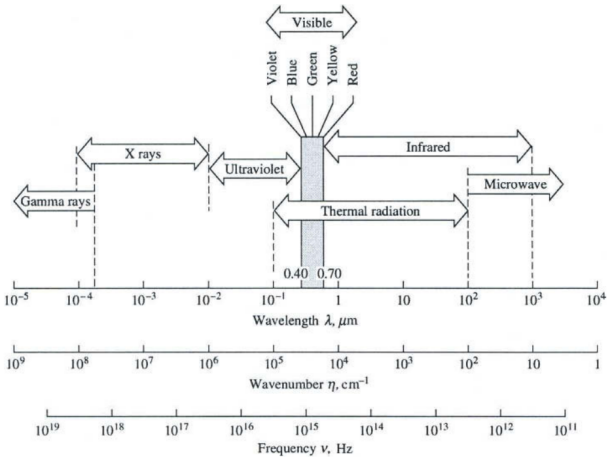


Figure 1. Electromagnetic spectrum - Credit MODEST, M.F. (1993). Radiative Heat Transfer. Academic Press.

For scientific purposes, infrared can be divided in three ranges of wavelength in which the application varies, see Table 1 .

Table 1. Wavelength bands in the infrared according to ISO 20473:2007

Band name	**wavelength**	**Uses ⟍ definition**
Near infrared (PIR, IR-A, NIR)	$0.7 - 3\mu$m	Reflected solar heat flux
Mid infrared (MIR, IR-B)	$3 - 50\mu$m	Thermal infrared
Far infrared (LIR, IR-C, FIR)	$50 - 1000\mu$m	Astronomy

Our work is concentrated in the mid infrared spectral band. Keep in mind that Table 1 represents the ISO 20473 division scheme, in the literature boundaries between bands can move slightly.
The Plank's law, proposed by Max Planck in 1901, allows to compute the black body emission spectrum for various temperatures (and only temperatures), see Figure 2 left. The black body is a theoretical construction, it represents perfect energy emitter at a given temperature, cf. Equation (20).

$$M^o_{\lambda,T} = \frac{C_1 \lambda^{-5}}{\exp^{\frac{C_2}{\lambda T}} - 1} \tag{25}$$

With λ the wavelength in m and T as the temperature in Kelvin. The C_1 and C_2 constants, respectively in W.m^2 and m.K are defined as follow:

$$\begin{aligned} C_1 &= 2hc^2\pi \\ C_2 &= h\frac{c}{k} \end{aligned} \tag{26}$$

with

- c, the electromagnetic wave speed (in vacuum c is the light speed in m.s^{-1}).
- $k = 1.381e^{-23}$ J.K^{-1} The Boltzmann (Entropy definition from Ludwig Boltzmann 1873). It can be seen as a proportionality factor between the temperature and the energy of a system.
- $h \approx 6,62606957e^{-34}$ J.s The Plank constant. It is the link between the photons energy and their frequency.

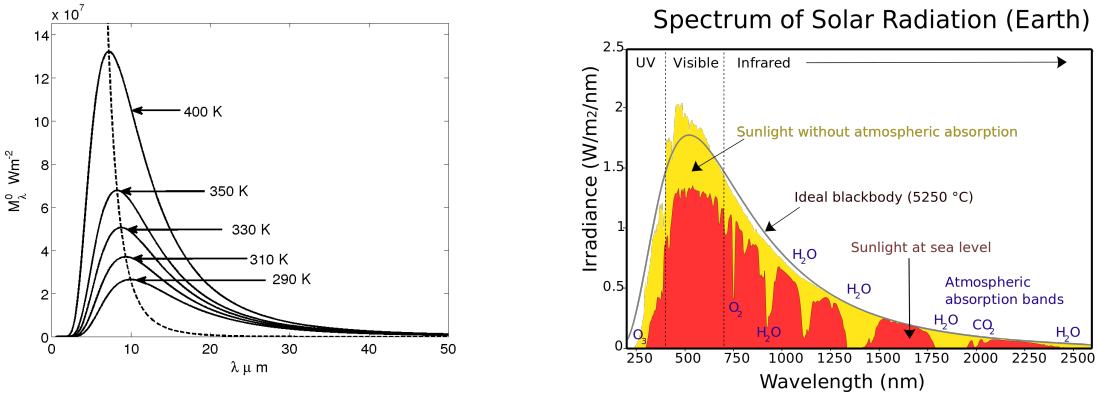


Figure 2. Left: Plank's law at various temperatures - Right: Energy spectrum of the atmosphere

By generalizing the Plank's law with the Stefan Boltzmann law (proposed first in 1879 and then in 1884 by Joseph Stefan and Ludwig Boltzmann), it is possible to address mathematically the energy spectrum of real body at each wavelength depending on the temperature, the optical condition and the real body properties, which is the base of the infrared thermography.

For example, Figure 2 right presents the energy spectrum of the atmosphere at various levels, it can be seen that the various properties of the atmosphere affect the spectrum at various wavelengths. Other important point is that the infrared solar heat flux can be approximated by a black body at 5523,15 K.

3.2.1.2. Infrared Thermography

The infrared thermography is a way to measure the thermal radiation received from a medium. With that information about the electromagnetic flux, it is possible to estimate the surface temperature of the body, see section 3.2.1.1 . Various types of detector can assure the measure of the electromagnetic radiation.

Those different detectors can take various forms and/or manufacturing process. For our research purposes, we use uncooled infrared camera using a matrix of microbolometers detectors. A microbolometer, as a lot of transducers, converts a radiation in electric current used to represent the physical quantity (here the heat flux).

This field of activity includes the use and the improvement of vision system, like in [3].

3.2.2. Heat transfer theory

Once the acquisition process is done, it is useful to model the heat conduction inside the cartesian domain Ω. Note that in opaque solid medium the heat conduction is the only mode of heat transfer. Proposed by Jean Baptiste Biot in 1804 and experimentally demonstrated by Joseph Fourier in 1821, the Fourier Law describes the heat flux inside a solid, cf Equation (22).

$$\varphi = k\nabla T \quad X \in \Omega \tag{27}$$

Where k is the thermal conductivity in W.m^{-1}.K ^{o}, ∇ is the gradient operator and φ is the heat flux density in Wm^{-2}. This law illustrates the first principle of thermodynamic (law of conservation of energy) and implies the second principle (irreversibility of the phenomenon). From this law it can be seen that the heat flux always goes from hot area to cold area.
An energy balance with respect to the first principle yields to the expression of the heat conduction in all point of the domain Ω, cf Equation (23). This equation has been proposed by Joseph Fourier in 1811.

$$\rho C \frac{\partial T(X,t)}{\partial t} = \nabla \cdot (k \nabla T) + P \quad X \in \Omega \tag{28}$$

With $\nabla.()$ the divergence operator, C the specific heat capacity in J.kg^{-1}.^{o}K^{-1}, ρ the volumetric mass density in kg. m^{-3}, X the space variable $X = \{x, y, z\}$ and P a possible internal heat production in W.m^{-3}.
To solve the system (23), it is necessary to express the boundaries conditions of the system. With the developments presented in section 3.2.1.1 and the Fourier's law, it is possible, for example, to express the thermal radiation and the convection phenomenon which can occur at $\partial\Omega$ the system boundaries, cf Equation (24).

$$\varphi = k \nabla T \cdot n = \underbrace{h \left(T_{fluid} - T_{Boundarie}\right)}_{\text{Convection}} + \underbrace{\epsilon \sigma_s \left(T^4_{environement} - T^4_{Boundary}\right)}_{\text{Radiation}} + \varphi_0 \quad X \in \partial\Omega \tag{29}$$

Equation (24) is the so called Robin condition on the boundary $\partial\Omega$, where n is the normal, h the convective heat transfer coefficient in W.m^{-2}.K^{-1} and φ_0 an external energy contribution W.m^{-2}, in cases where the external energy contribution is artificial and controlled we call it active thermography (spotlight etc...), otherwise it is called passive thermography (direct solar heat flux).
The systems presented in the different sections above (3.2.1 to 3.2.2) are useful to build physical models in order to represents the measured quantity. To estimate key parameters, as the conductivity, model inversion is used, the next section will introduce that principle.

3.2.3. Inverse model for parameters estimation

Lets take any model A which can for example represent the conductive heat transfer in a medium, the model is solved for a parameter vector P and it yields another vector b, cf Equation (25). For example if A represents the heat transfer, b can be the temperature evolution.

$$AP = b \tag{30}$$

With A a matrix of size $n \times m$, P a vector of size m and b of size n, preferentially $n >> P$. This model is called direct model, the inverse model consist to find a vector P which satisfy the results b of the direct model. For that we need to inverse the matrix A, cf Equation (26).

$$P = A^{-1} b \tag{31}$$

Here we want to find the solution AP which is closest to the acquired measures M, Equation (27).

$$AP \approx \mathcal{M} \tag{32}$$

To do that it is important to respect the well posed condition established by Jacques Hadamard in 1902

- A solution exists.
- The solution is unique.
- The solution's behavior changes continuously with the initial conditions.

Unfortunately those condition are rarely respected in our field of study. That is why we dont solve directly the system (27) but we minimise the quadratic coast function (28) which represents the Legendre-Gauss least square algorithm for linear problems.

$$min_P \left(\|AP - \mathcal{M}\|^2 \right) = min_P \left(\mathcal{F} \right) \tag{33}$$

Where $\mathcal{F}$ can be a product of matrix.

$$\mathcal{F} = \left[AP - \mathcal{M} \right]^T \left[AP - \mathcal{M} \right]$$

In some cases the problem is still ill-posed and need to be regularized for example using the Tikhonov regularization. An elegant way to minimize the cost function $\mathcal{F}$ is compute the gradient, Equation (29) and find where it is equal to zero.

$$\nabla \mathcal{F}(P) = 2 \left[-\frac{\partial AP^T}{\partial P} \right] \left[AP - \mathcal{M} \right] = 2J(P)^T \left[AP - \mathcal{M} \right] \tag{34}$$

Where J is the sensitivity matrix of the model A with respect to the parameter vector P.
Until now the inverse method proposed is valid only when the model A is linearly dependent of its parameter P, for the heat equation it is the case when the external heat flux has to be estimated, φ_0 in Equation (24). For all the other parameters, like the conductivity k the model is non-linearly dependant of its parameter P. For such case the use of iterative algorithm is needed, for example the Levenberg-Marquardt algorithm, cf Equation (30).

$$P^{k+1} = P^k + \left[\left(J^k \right)^T J^k + \mu^k \Omega^k \right]^{-1} \left(J^k \right)^T \left[\mathcal{M} - A(P^k) \right] \tag{35}$$

Equation (30) is solved iteratively at each loop k. Some of our results with such linear or non linear method can be seen in [4] or [2], more specifically [1] is a custom implementation of the Levenberg-Marquardt algorithm based on the adjoint method (developed by Jacques Louis Lions in 1968) coupled to the conjugate gradient algorithm to estimate wide properties field in a medium.

3.3. Reflectometry-based methods for electrical engineering and for civil engineering

The fast development of electronic devices in modern engineering systems involves more and more connections through cables, and consequently, with an increasing number of connection failures. Wires and connectors are subject to ageing and degradation, sometimes under severe environmental conditions. In many applications, the reliability of electrical connexions is related to the quality of production or service, whereas in critical applications reliability becomes also a safety issue. It is thus important to design smart diagnosis systems able to detect connection defects in real time. This fact has motivated research projects on methods for fault diagnosis in this field. Some of these projects are based on techniques of reflectometry, which consist in injecting waves into a cable or a network and in analyzing the reflections. Depending on the injected waveforms and on the methods of analysis, various techniques of reflectometry are available. They all have the common advantage of being non destructive.

At Inria the research activities on reflectometry started within the SISYPHE EPI several years ago and now continue in the I4S EPI. Our most notable contribution in this area is a method based on the *inverse scattering* theory for the computation of *distributed characteristic impedance* along a cable from reflectometry measurements [14], [11], [50]. It provides an efficient solution for the diagnosis of *soft* faults in electrical cables, like in the example illustrated in Figure 3 . While most reflectometry methods for fault diagnosis are based on the detection and localization of impedance discontinuity, our method yielding the spatial profile of the characteristic impedance is particularly suitable for the diagnosis of soft faults *with no or weak impedance discontinuities*.

Fault diagnosis for wired networks have also been studied in Inria [52], [48]. The main results concern, on the one hand, simple star-shaped networks from measurements made at a single node, on the other hand, complex networks of arbitrary topological structure with complete node observations.

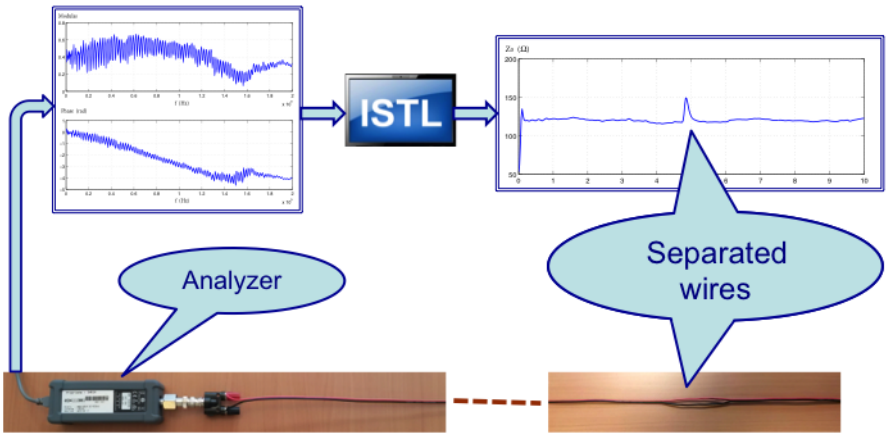


Figure 3. Inverse scattering software (ISTL) for cable soft fault diagnosis.

Though initially our studies on reflectometry were aiming at applications in electrical engineering, since the creation of the I4S team, we are also investigating applications in the field of civil engineering, by using electrical cables as sensors for monitoring changes in mechanical structures.

What follows is about some basic elements on mathematical equations of electric cables and networks, the main approach we follow in our study, and our future research directions.

3.3.1. Mathematical model of electric cables and networks

A cable excited by a signal generator can be characterized by the telegrapher's equations [49]

$$\begin{aligned} &\frac{\partial}{\partial z}V(t,z) + L(z)\frac{\partial}{\partial t}I(t,z) + R(z)I(t,z) = 0 \\ &\frac{\partial}{\partial z}I(t,z) + C(z)\frac{\partial}{\partial t}V(t,z) + G(z)V(t,z) = 0 \end{aligned} \tag{36}$$

where t represents the time, z is the longitudinal coordinate along the cable, $V(t,z)$ and $I(t,z)$ are respectively the voltage and the current in the cable at the time instant t and at the position z, $R(z), L(z), C(z)$ and $G(z)$ denote respectively the series resistance, the inductance, the capacitance and the shunt conductance per unit length of the cable at the position z. The left end of the cable (corresponding to $z = a$) is connected to a voltage source $V_s(t)$ with internal impedance R_s. The quantities $V_s(t)$, R_s, $V(t,a)$ and $I(t,a)$ are related by

$$V(t,a) = V_s(t) - R_s I(t,a). \tag{37}$$

At the right end of the cable (corresponding to $z = b$), the cable is connected to a load of impedance R_L, such that

$$V(t,b) \quad = R_L I(t,b). \tag{38}$$

One way for deriving the above model is to spatially discretize the cable and to characterize each small segment with 4 basic lumped parameter elements for the j-th segment: a resistance ΔR_j, an inductance ΔL_j, a capacitance ΔC_j and a conductance ΔG_j. The entire circuit is described by a system of ordinary differential equations. When the spatial discretization step size tends to zero, the limiting model leads to the telegrapher's equations.

A wired network is a set of cables connected at some nodes, where loads and sources can also be connected. Within each cable the current and voltage satisfy the telegrapher's equations, whereas at each node the current and voltage satisfy the Kirchhoff's laws, unless in case of connector failures.

3.3.2. The inverse scattering theory applied to cables

The inverse scattering transform was developed during the 1970s-1980s for the analysis of some nonlinear partial differential equations [47]. The visionary idea of applying this theory to solving the cable inverse problem goes also back to the 1980s [46]. After having completed some theoretic results directly linked to practice [14], [50], we started to successfully apply the inverse scattering theory to cable soft fault diagnosis, in collaboration with GEEPS-SUPELEC [11].

To link electric cables to the inverse scattering theory, the telegrapher's equations are transformed in a few steps to fit into a particular form studied in the inverse scattering theory. The Fourier transform is first applied to obtain a frequency domain model, the spatial coordinate z is then replaced by the propagation time

$$x(z) = \int_0^z \sqrt{L(s)C(s)}ds$$

and the frequency domain variables $V(\omega,x), I(\omega,x)$ are replaced by the pair

$$\begin{aligned} \nu_1(\omega,x) &= \frac{1}{2}\left[Z_0^{-\frac{1}{2}}(x)U(\omega,x) - Z_0^{\frac{1}{2}}(x)I(\omega,x)\right] \\ \nu_2(\omega,x) &= \frac{1}{2}\left[Z_0^{-\frac{1}{2}}(x)U(\omega,x) + Z_0^{\frac{1}{2}}(x)I(\omega,x)\right] \end{aligned} \tag{39}$$

with

$$Z_0(x) = \sqrt{\frac{L(x)}{C(x)}}. \tag{40}$$

These transformations lead to the Zakharov-Shabat equations

$$\begin{aligned} \frac{d\nu_1(\omega,x)}{dx} + ik\nu_1(\omega,x) &= q^*(x)\nu_1(\omega,x) + q^+(x)\nu_2(\omega,x) \\ \frac{d\nu_2(\omega,x)}{dx} - ik\nu_2(\omega,x) &= q^-(x)\nu_1(\omega,x) - q^*(x)\nu_2(\omega,x) \end{aligned} \tag{41}$$

with

$$\begin{aligned} q^{\pm}(x) &= -\frac{1}{4}\frac{d}{dx}\left[ln\frac{L(x)}{C(x)}\right] \mp \frac{1}{2}\left[\frac{R(x)}{L(x)} - \frac{G(x)}{C(x)}\right] \\ &= -\frac{1}{2Z_0(x)}\frac{d}{dx}Z_0(x) \mp \frac{1}{2}\left[\frac{R(x)}{L(x)} - \frac{G(x)}{C(x)}\right] \\ q^{*}(x) &= \frac{1}{2}\left[\frac{R(x)}{L(x)} + \frac{G(x)}{C(x)}\right]. \end{aligned} \tag{42}$$

These equations have been well studied in the inverse scattering theory, for the purpose of determining partly the "potential functions" $q^{\pm}(x)$ and $q^{*}(x)$ from the scattering data matrix, which turns out to correspond to the data typically collected with reflectometry instruments. For instance, it is possible to compute the function $Z_0(x)$ defined in (35), often known as the characteristic impedance, from the reflection coefficient measured at one end of the cable. Such an example is illustrated in Figure 3 . Any fault affecting the characteristic impedance, like in the example of Figure 3 caused by a slight geometric deformation, can thus be efficiently detected, localized and characterized.

3.4. Research Program

The research will first focus on the extension and implementation of current techniques as developed in I4S and IFSTTAR. Before doing any temperature rejection on large scale structures as planned, we need to develop good and accurate models of thermal fields. We also need to develop robust and efficient versions of our algorithms, mainly the subspace algorithms before envisioning linking them with physical models. Briefly, we need to mature our statistical toolset as well as our physical modeling before mixing them together later on.

3.4.1. Vibration analysis and monitoring

3.4.1.1. Direct vibration modeling under temperature changes

This task builds upon what has been achieved in the CONSTRUCTIF project, where a simple formulation of the temperature effect has been exhibited, based on relatively simple assumptions. The next step is to generalize this modeling to a realistic large structure under complex thermal changes. Practically, temperature and resulting structural prestress and pre strains of thermal origin are not uniform and civil structures are complex. This leads to a fully 3D temperature field, not just a single value. Inertia effects also forbid a trivial prediction of the temperature based on current sensor outputs while ignoring past data. On the other side, the temperature is seen as a nuisance. That implies that any damage detection procedure has first to correct the temperature effect prior to any detection.

Modeling vibrations of structures under thermal prestress does and will play an important role in the static correction of kinematic measurements, in health monitoring methods based on vibration analysis as well as in durability and in the active or semi-active control of civil structures that by nature are operated under changing environmental conditions. As a matter of fact, using temperature and dynamic models the project aims at correcting the current vibration state from induced temperature effects, such that damage detection algorithms rely on a comparison of this thermally corrected current vibration state with a reference state computed or measured at a reference temperature. This approach is expected to cure damage detection algorithms from the environmental variations.

I4S will explore various ways of implementing this concept, notably within the FUI SIPRIS project.

3.4.1.2. Damage localization algorithms (in the case of localized damages such as cracks)

During the CONSTRUCTIF project, both feasibility and efficiency of some damage detection and localization algorithms were proved. Those methods are based on the tight coupling of statistical algorithms with finite element models. It has been shown that effective localization of some damaged elements was possible, and this was validated on a numerical simulated bridge deck model. Still, this approach has to be validated on real structures.

On the other side, new localization algorithms are currently investigated such as the one developed conjointly with University of Boston and tested within the framework of FP7 ISMS project. These algorithms will be implemented and tested on the PEGASE platform as well as all our toolset.

When possible, link with temperature rejection will be done along the lines of what has been achieved in the CONSTRUCTIF project.

3.4.1.3. Uncertainty quantification for system identification algorithms

Some emphasis will be put on expressing confidence intervals for system identification. It is a primary goal to take into account the uncertainty within the identification procedure, using either identification algorithms derivations or damage detection principles. Such algorithms are critical for both civil and aeronautical structures monitoring. It has been shown that confidence intervals for estimation parameters can theoretically be related to the damage detection techniques and should be computed as a function of the Fisher information matrix associated to the damage detection test. Based on those assumptions, it should be possible to obtain confidence intervals for a large class of estimates, from damping to finite elements models. Uncertainty considerations are also deeply investigated in collaboration with Dassault Aviation in Mellinger PhD thesis or with Northeastern University, Boston, within Gallegos PhD thesis.

3.4.2. Reflectometry-based methods for civil engineering structure health monitoring

The inverse scattering method we developed is efficient for the diagnosis of all soft faults affecting the characteristic impedance, the major parameter of a cable. In some particular applications, however, faults would rather affect the series resistance (ohmic loss) or shunt conductance (leakage loss) than the characteristic impedance. The first method we developed for the diagnosis of such losses had some numerical stability problems. The new method is much more reliable and efficient. It is also important to develop efficient solutions for long cables, up to a few kilometers.

For wired networks, the methods we already developed cover either the case of simple networks with a single node measurement or the case of complex networks with complete node measurements. Further developments are still necessary for intermediate situations.

In terms of applications, the use of electric cables as sensors for the monitoring of various structures is still at its beginning. We believe that this new technology has a strong potential in different fields, notably in civil engineering and in materials engineering.

3.4.3. Non Destructive testing of CFRP bonded on concrete through active thermography

Strengthening or retrofitting of reinforced concrete structures by externally bonded fiber-reinforced polymer (FRP) systems is now a commonly accepted and widespread technique. However, the use of bonding techniques always implies following rigorous installation procedures. The number of carbon fiber-reinforced polymer (CFRP) sheets and the glue layer thickness are designed by civil engineers to address strengthening objectives. Moreover, professional crews have to be trained accordingly in order to ensure the durability and long-term performance of the FRP reinforcements. Conformity checking through an ‘in situ’ verification of the bonded FRP systems is then highly desirable. The quality control programme should involve a set of adequate inspections and tests. Visual inspection and acoustic sounding (hammer tap) are commonly used to detect delaminations (disbonds). Nevertheless, these techniques are unable to provide sufficient information about the depth (in case of multilayered composite) and width of the disbonded areas. They are also incapable of evaluating the degree of adhesion between the FRP and the substrate (partial delamination, damage of the resin and poor mechanical properties of the resin). Consequently, rapid and efficient inspection methods are required. Among the non-destructive (NDT) methods currently under study, active infrared thermography is investigated due to its ability to be used in the field. In such context and to reach the aim of having an in situ efficient NDT method, we carried out experiments and subsequent data analysis using thermal excitation. Image processing, inverse thermal modelling and 3D numerical simulations are used and then applied to experimental data obtained in laboratory conditions.

3.4.4. IRSHM: Multi-Sensing system for outdoor thermal monitoring

Ageing of transport infrastructures combined with traffic and climatic solicitations contribute to the reduction of their performances. To address and quantify the resilience of civil engineering structure, investigations on robust, fast and efficient methods are required. Among research works carried out at IFSTTAR, methods for long term monitoring face an increasing demand. Such works take benefits of this last decade technological progresses in ICT domain.

Thanks to IFSTTAR years of experience in large scale civil engineering experiment, I4S is able to perform very long term thermal monitoring of structures exposed to environmental condition, as the solar heat flux, natural convection or seasonal perturbation. Informations system are developed to asses the data acquisition and researchers work on the quantification of the data to detect flaws emergence on structure, those techniques are also used to diagnose thermal insulation of buildings or monitoring of guided transport infrastructures, Figure 4 left. Experiments are carried out on a real transport infrastructure open to traffic and buildings. The detection of the inner structure of the deck is achieved by image processing techniques (as FFT), principal component thermography (PCT), Figure 4 right, or characterization of the inner structure thanks to an original image processing approach.

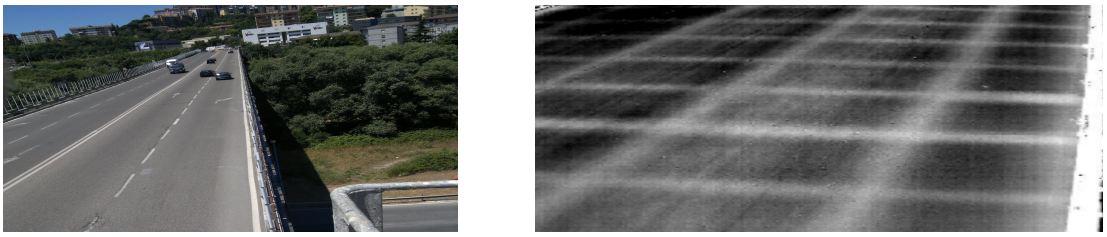


Figure 4. Left: Image in the visible spectrum of the deck surface - Right: PCT result on a bridge deck

For the next few years, I4S is actively implied in the SenseCity EQUIPEX (http://sense-city.ifsttar.fr/) where our informations systems are used to monitor a mini-city replica, Figure 5 .

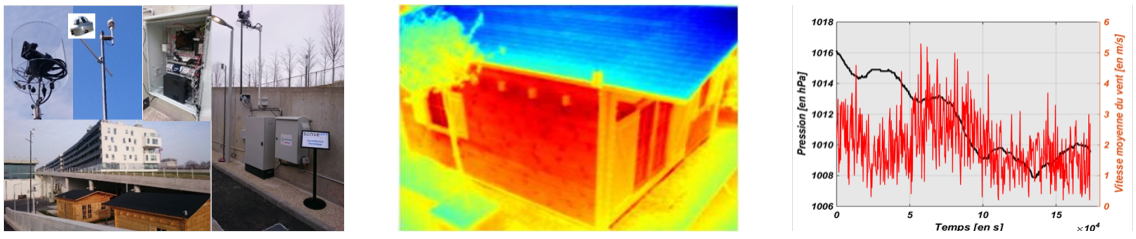


Figure 5. Various view and results of the SenseCity experimentation site - (site and hardware view, IR imaging, Environmental Monitoring)

3.4.5. R5G: The 5th Generation Road

The road has to reinvent itself periodically in response to innovations, societal issues and rising user expectations. The 5th Generation Road (R5G) focuses firmly on the future and sets out to be automated, safe, sustainable and suited to travel needs. Several research teams are involved in work related to this flagship

project for IFSTTAR, which is a stakeholder in the Forever Open Road. Through its partnership with the COSYS (IFSTTAR) department, I4S is fully involved in the development of the 5th Generation Road.

Most of the innovations featured in R5G are now mature, for example communication and few solutions for energy exchange between the infrastructure, the vehicle and the network manager; recyclable materials with the potential for self-diagnosis and repair, a pavement surface that remains permanently optimal irrespective of climatic variations… Nevertheless, implementing them on an industrial scale at a reasonable cost still represents a real challenge. Consultation with the stakeholders (researchers, industry, road network owners and users) has already established the priorities for the creation of full-scale demonstrators. The next stages are to achieve synergy between the technologies tested by the demonstrators, to manage the interfaces and get society to adopt R5G.

INOCS Project-Team

3. Research Program

3.1. Introduction

An optimization problem consists in finding a best solution from a set of feasible solutions. Such a problem can be typically modeled as a mathematical program in which decision variables must:

1. satisfy a set of constraints that translate the feasibility of the solution and
2. optimize some (or several) objective function(s). Optimization problems are usually classified according to types of decision to be taken into strategic, tactical and operational problems.

We consider that an optimization problem presents a complex structure when it involves decisions of different types/nature (i.e. strategic, tactical or operational), and/or presenting some hierarchical leader-follower structure. The set of constraints may usually be partitioned into global constraints linking variables associated with the different types/nature of decision and constraints involving each type of variables separately. Optimization problems with a complex structure lead to extremely challenging problems since a global optimum with respect to the whole sets of decision variables and of constraints must be determined.

Significant progresses have been made in optimization to solve academic problems. Nowadays large-scale instances of some NP-Hard problems are routinely solved to optimality. *Our vision within INOCS is to make the same advances while addressing CS optimization problems*. To achieve this goal we aim to develop global solution approaches at the opposite of the current trend. INOCS team members have already proposed some successful methods following this research lines to model and solve CS problems (e.g. ANR project RESPET, Brotcorne *et al.* 2011, 2012, Gendron *et al.* 2009, Strack *et al.* 2009). However, these are preliminary attempts and a number of challenges regarding modeling and methodological issues have still to be met.

3.2. Modeling problems with complex structures

A classical optimization problem can be formulated as follows:

$$\begin{array}{ll} \min & f(x) \\ s.t. & x \in X. \end{array} \tag{43}$$

In this problem, X is the set of feasible solutions. Typically, in mathematical programming, X is defined by a set of constraints. x may be also limited to non-negative integer values.

INOCS team plan to address optimization problem where two types of decision are addressed jointly and are interrelated. More precisely, let us assume that variables x and y are associated with these decisions. A generic model for CS problems is the following:

$$\begin{array}{ll} \min & g(x,y) \\ s.t. & x \in X, \\ & (x,y) \in XY, \\ & y \in Y(x). \end{array} \tag{44}$$

In this model, X is the set of feasible values for x. XY is the set of feasible values for x and y jointly. This set is typically modeled through linking constraints. Last, $Y(x)$ is the set of feasible values for y for a given x. In INOCS, we do not assume that $Y(x)$ has any properties.

The INOCS team plans to model optimization CS problems according to three types of optimization paradigms: large scale complex structures optimization, bilevel optimization and robust/stochastic optimization. These paradigms instantiate specific variants of the generic model.

Large scale complex structures optimization problems can be formulated through the simplest variant of the generic model given above. In this case, it is assumed that $Y(x)$ does not depend on x. In such models, X and Y are associated with constraints on x and on y, XY are the linking constraints. x and y can take continuous or integer values. Note that all the problem data are deterministically known.

Bilevel programs allow the modeling of situations in which a decision-maker, hereafter the leader, optimizes his objective by taking explicitly into account the response of another decision maker or set of decision makers (the follower) to his/her decisions. Bilevel programs are closely related to Stackelberg (leader-follower) games as well as to the principal-agent paradigm in economics. In other words, bilevel programs can be considered as demand-offer equilibrium models where the demand is the result of another mathematical problem. Bilevel problems can be formulated through the generic CS model when $Y(x)$ corresponds to the optimal solutions of a mathematical program defined for a given x, i.e. $Y(x) = \arg\min \{h(x,y)|y \in Y_2, (x,y) \in XY_2\}$ where Y_2 is defined by a set of constraints on y, and XY_2 is associated with the linking constraints.

In robust/stochastic optimization, it is assumed that the data related to a problem are subject to uncertainty. In stochastic optimization, probability distributions governing the data are known, and the objective function involves mathematical expectation(s). In robust optimization, uncertain data take value within specified sets, and the function to optimize is formulated in terms of a min-max objective typically (the solution must be optimal for the worst-case scenario). A standard modeling of uncertainty on data is obtained by defining a set of possible scenarios that can be described explicitly or implicitly. In stochastic optimization, in addition, a probability of occurrence is associated with each scenario and the expected objective value is optimized.

3.3. Solving problems with complex structures

Standard solution methods developed for CS problems solve independent sub-problems associated with each type of variables without explicitly integrating their interactions or integrating them iteratively in a heuristic way. However these subproblems are intrinsically linked and should be addressed jointly. In *mathematicaloptimization* a classical approach is to approximate the convex hull of the integer solutions of the model by its linear relaxation. The main solution methods are (1) polyhedral solution methods which strengthen this linear relaxation by adding valid inequalities, (2) decomposition solution methods (Dantzig Wolfe, Lagrangian Relaxation, Benders decomposition) which aim to obtain a better approximation and solve it by generating extreme points/rays. Main challenges are (1) the analysis of the strength of the cuts and their separations for polyhedral solution methods, (2) the decomposition schemes and (3) the extreme points/rays generations for the decomposition solution methods.

The main difficulty in solving *bilevel problems* is due to their non convexity and non differentiability. Even linear bilevel programs, where all functions involved are affine, are computationally challenging despite their apparent simplicity. Up to now, much research has been devoted to bilevel problems with linear or convex follower problems. In this case, the problem can be reformulated as a single-level program involving complementarity constraints, exemplifying the dual nature, continuous and combinatorial, of bilevel programs.

MATHERIALS Project-Team

3. Research Program

3.1. Research Program

Our group, originally only involved in electronic structure computations, continues to focus on many numerical issues in quantum chemistry, but now expands its expertise to cover several related problems at larger scales, such as molecular dynamics problems and multiscale problems. The mathematical derivation of continuum energies from quantum chemistry models is one instance of a long-term theoretical endeavour.

3.1.1. Electronic structure of large systems

Quantum Chemistry aims at understanding the properties of matter through the modelling of its behavior at a subatomic scale, where matter is described as an assembly of nuclei and electrons. At this scale, the equation that rules the interactions between these constitutive elements is the Schrödinger equation. It can be considered (except in few special cases notably those involving relativistic phenomena or nuclear reactions) as a universal model for at least three reasons. First it contains all the physical information of the system under consideration so that any of the properties of this system can in theory be deduced from the Schrödinger equation associated to it. Second, the Schrödinger equation does not involve any empirical parameters, except some fundamental constants of Physics (the Planck constant, the mass and charge of the electron, ...); it can thus be written for any kind of molecular system provided its chemical composition, in terms of natures of nuclei and number of electrons, is known. Third, this model enjoys remarkable predictive capabilities, as confirmed by comparisons with a large amount of experimental data of various types. On the other hand, using this high quality model requires working with space and time scales which are both very tiny: the typical size of the electronic cloud of an isolated atom is the Angström (10^{-10} meters), and the size of the nucleus embedded in it is 10^{-15} meters; the typical vibration period of a molecular bond is the femtosecond (10^{-15} seconds), and the characteristic relaxation time for an electron is 10^{-18} seconds. Consequently, Quantum Chemistry calculations concern very short time (say 10^{-12} seconds) behaviors of very small size (say 10^{-27} m^3) systems. The underlying question is therefore whether information on phenomena at these scales is useful in understanding or, better, predicting macroscopic properties of matter. It is certainly not true that *all* macroscopic properties can be simply upscaled from the consideration of the short time behavior of a tiny sample of matter. Many of them derive from ensemble or bulk effects, that are far from being easy to understand and to model. Striking examples are found in solid state materials or biological systems. Cleavage, the ability of minerals to naturally split along crystal surfaces (e.g. mica yields to thin flakes), is an ensemble effect. Protein folding is also an ensemble effect that originates from the presence of the surrounding medium; it is responsible for peculiar properties (e.g. unexpected acidity of some reactive site enhanced by special interactions) upon which vital processes are based. However, it is undoubtedly true that *many* macroscopic phenomena originate from elementary processes which take place at the atomic scale. Let us mention for instance the fact that the elastic constants of a perfect crystal or the color of a chemical compound (which is related to the wavelengths absorbed or emitted during optic transitions between electronic levels) can be evaluated by atomic scale calculations. In the same fashion, the lubricative properties of graphite are essentially due to a phenomenon which can be entirely modeled at the atomic scale. It is therefore reasonable to simulate the behavior of matter at the atomic scale in order to understand what is going on at the macroscopic one. The journey is however a long one. Starting from the basic principles of Quantum Mechanics to model the matter at the subatomic scale, one finally uses statistical mechanics to reach the macroscopic scale. It is often necessary to rely on intermediate steps to deal with phenomena which take place on various *mesoscales*. It may then be possible to couple one description of the system with some others within the so-called *multiscale* models. The sequel indicates how this journey can be completed focusing on the first smallest scales (the subatomic one), rather than on the larger ones. It has already been mentioned that at the subatomic scale, the behavior of nuclei and electrons is governed by the Schrödinger equation, either in its time-dependent form or in its time-independent form. Let us only mention at this point that

- both equations involve the quantum Hamiltonian of the molecular system under consideration; from a mathematical viewpoint, it is a self-adjoint operator on some Hilbert space; *both* the Hilbert space and the Hamiltonian operator depend on the nature of the system;
- also present into these equations is the wavefunction of the system; it completely describes its state; its L^2 norm is set to one.

The time-dependent equation is a first-order linear evolution equation, whereas the time-independent equation is a linear eigenvalue equation. For the reader more familiar with numerical analysis than with quantum mechanics, the linear nature of the problems stated above may look auspicious. What makes the numerical simulation of these equations extremely difficult is essentially the huge size of the Hilbert space: indeed, this space is roughly some symmetry-constrained subspace of $L^2(\mathbb{R}^d)$, with $d = 3(M + N)$, M and N respectively denoting the number of nuclei and the number of electrons the system is made of. The parameter d is already 39 for a single water molecule and rapidly reaches 10^6 for polymers or biological molecules. In addition, a consequence of the universality of the model is that one has to deal at the same time with several energy scales. In molecular systems, the basic elementary interaction between nuclei and electrons (the two-body Coulomb interaction) appears in various complex physical and chemical phenomena whose characteristic energies cover several orders of magnitude: the binding energy of core electrons in heavy atoms is 10^4 times as large as a typical covalent bond energy, which is itself around 20 times as large as the energy of a hydrogen bond. High precision or at least controlled error cancellations are thus required to reach chemical accuracy when starting from the Schrödinger equation. Clever approximations of the Schrödinger problems are therefore needed. The main two approximation strategies, namely the Born-Oppenheimer-Hartree-Fock and the Born-Oppenheimer-Kohn-Sham strategies, end up with large systems of coupled *nonlinear* partial differential equations, each of these equations being posed on $L^2(\mathbb{R}^3)$. The size of the underlying functional space is thus reduced at the cost of a dramatic increase of the mathematical complexity of the problem: nonlinearity. The mathematical and numerical analysis of the resulting models has been the major concern of the project-team for a long time. In the recent years, while part of the activity still follows this path, the focus has progressively shifted to problems at other scales.

As the size of the systems one wants to study increases, more efficient numerical techniques need to be resorted to. In computational chemistry, the typical scaling law for the complexity of computations with respect to the size of the system under study is N^3, N being for instance the number of electrons. The Holy Grail in this respect is to reach a linear scaling, so as to make possible simulations of systems of practical interest in biology or material science. Efforts in this direction must address a large variety of questions such as

- how can one improve the nonlinear iterations that are the basis of any *ab initio* models for computational chemistry?
- how can one more efficiently solve the inner loop which most often consists in the solution procedure for the linear problem (with frozen nonlinearity)?
- how can one design a sufficiently small variational space, whose dimension is kept limited while the size of the system increases?

An alternative strategy to reduce the complexity of *ab initio* computations is to try to couple different models at different scales. Such a mixed strategy can be either a sequential one or a parallel one, in the sense that

- in the former, the results of the model at the lower scale are simply used to evaluate some parameters that are inserted in the model for the larger scale: one example is the parameterized classical molecular dynamics, which makes use of force fields that are fitted to calculations at the quantum level;
- while in the latter, the model at the lower scale is concurrently coupled to the model at the larger scale: an instance of such a strategy is the so called QM/MM coupling (standing for Quantum Mechanics/Molecular Mechanics coupling) where some part of the system (typically the reactive site of a protein) is modeled with quantum models, that therefore accounts for the change in the electronic structure and for the modification of chemical bonds, while the rest of the system (typically the inert part of a protein) is coarse grained and more crudely modeled by classical mechanics.

The coupling of different scales can even go up to the macroscopic scale, with methods that couple a microscopic representation of matter, or at least a mesoscopic one, with the equations of continuum mechanics at the macroscopic level.

3.1.2. Computational Statistical Mechanics

The orders of magnitude used in the microscopic representation of matter are far from the orders of magnitude of the macroscopic quantities we are used to: The number of particles under consideration in a macroscopic sample of material is of the order of the Avogadro number $\mathcal{N}_A \sim 6 \times 10^{23}$, the typical distances are expressed in Å (10^{-10} m), the energies are of the order of $k_{\mathrm{B}}T \simeq 4 \times 10^{-21}$ J at room temperature, and the typical times are of the order of 10^{-15} s.

To give some insight into such a large number of particles contained in a macroscopic sample, it is helpful to compute the number of moles of water on earth. Recall that one mole of water corresponds to 18 mL, so that a standard glass of water contains roughly 10 moles, and a typical bathtub contains 10^5 mol. On the other hand, there are approximately 10^{18} m^3 of water in the oceans, *i.e.*7×10^{22} mol, a number comparable to the Avogadro number. This means that inferring the macroscopic behavior of physical systems described at the microscopic level by the dynamics of several millions of particles only is like inferring the ocean's dynamics from hydrodynamics in a bathtub...

For practical numerical computations of matter at the microscopic level, following the dynamics of every atom would require simulating $\mathcal{N}_A$ atoms and performing $\mathrm{O}(10^{15})$ time integration steps, which is of course impossible! These numbers should be compared with the current orders of magnitude of the problems that can be tackled with classical molecular simulation, where several millions of atoms only can be followed over time scales of the order of a few microseconds.

Describing the macroscopic behavior of matter knowing its microscopic description therefore seems out of reach. Statistical physics allows us to bridge the gap between microscopic and macroscopic descriptions of matter, at least on a conceptual level. The question is whether the estimated quantities for a system of N particles correctly approximate the macroscopic property, formally obtained in the thermodynamic limit $N \to +\infty$ (the density being kept fixed). In some cases, in particular for simple homogeneous systems, the macroscopic behavior is well approximated from small-scale simulations. However, the convergence of the estimated quantities as a function of the number of particles involved in the simulation should be checked in all cases.

Despite its intrinsic limitations on spatial and timescales, molecular simulation has been used and developed over the past 50 years, and its number of users keeps increasing. As we understand it, it has two major aims nowadays.

First, it can be used as a *numerical microscope*, which allows us to perform "computer" experiments. This was the initial motivation for simulations at the microscopic level: physical theories were tested on computers. This use of molecular simulation is particularly clear in its historic development, which was triggered and sustained by the physics of simple liquids. Indeed, there was no good analytical theory for these systems, and the observation of computer trajectories was very helpful to guide the physicists' intuition about what was happening in the system, for instance the mechanisms leading to molecular diffusion. In particular, the pioneering works on Monte-Carlo methods by Metropolis *et al.*, and the first molecular dynamics simulation of Alder and Wainwright were performed because of such motivations. Today, understanding the behavior of matter at the microscopic level can still be difficult from an experimental viewpoint (because of the high resolution required, both in time and in space), or because we simply do not know what to look for! Numerical simulations are then a valuable tool to test some ideas or obtain some data to process and analyze in order to help assessing experimental setups. This is particularly true for current nanoscale systems.

Another major aim of molecular simulation, maybe even more important than the previous one, is to compute macroscopic quantities or thermodynamic properties, typically through averages of some functionals of the system. In this case, molecular simulation is a way to obtain *quantitative* information on a system, instead of resorting to approximate theories, constructed for simplified models, and giving only qualitative answers. Sometimes, these properties are accessible through experiments, but in some cases only numerical

computations are possible since experiments may be unfeasible or too costly (for instance, when high pressure or large temperature regimes are considered, or when studying materials not yet synthesized). More generally, molecular simulation is a tool to explore the links between the microscopic and macroscopic properties of a material, allowing one to address modelling questions such as "Which microscopic ingredients are necessary (and which are not) to observe a given macroscopic behavior?"

3.1.3. Homogenization and related problems

Over the years, the project-team has developed an increasing expertise on how to couple models written at the atomistic scale with more macroscopic models, and, more generally, an expertise in multiscale modelling for materials science.

The following observation motivates the idea of coupling atomistic and continuum representation of materials. In many situations of interest (crack propagation, presence of defects in the atomistic lattice, ...), using a model based on continuum mechanics is difficult. Indeed, such a model is based on a macroscopic constitutive law, the derivation of which requires a deep qualitative and quantitative understanding of the physical and mechanical properties of the solid under consideration. For many solids, reaching such an understanding is a challenge, as loads they are subjected to become larger and more diverse, and as experimental observations helping designing such models are not always possible (think of materials used in the nuclear industry). Using an atomistic model in the whole domain is not possible either, due to its prohibitive computational cost. Recall indeed that a macroscopic sample of matter contains a number of atoms on the order of 10^{23}. However, it turns out that, in many situations of interest, the deformation that we are looking for is not smooth in *only a small part* of the solid. So, a natural idea is to try to take advantage of both models, the continuum mechanics one and the atomistic one, and to couple them, in a domain decomposition spirit. In most of the domain, the deformation is expected to be smooth, and reliable continuum mechanics models are then available. In the rest of the domain, the expected deformation is singular, so that one needs an atomistic model to describe it properly, the cost of which remains however limited as this region is small.

From a mathematical viewpoint, the question is to couple a discrete model with a model described by PDEs. This raises many questions, both from the theoretical and numerical viewpoints:

- first, one needs to derive, from an atomistic model, continuum mechanics models, under some regularity assumptions that encode the fact that the situation is smooth enough for such a macroscopic model to provide a good description of the materials;
- second, couple these two models, e.g. in a domain decomposition spirit, with the specificity that models in both domains are written in a different language, that there is no natural way to write boundary conditions coupling these two models, and that one would like the decomposition to be self-adaptive.

More generally, the presence of numerous length scales in material science problems represents a challenge for numerical simulation, especially when some *randomness* is assumed on the materials. It can take various forms, and includes defects in crystals, thermal fluctuations, and impurities or heterogeneities in continuous media. Standard methods available in the literature to handle such problems often lead to very costly computations. Our goal is to develop numerical methods that are more affordable. Because we cannot embrace all difficulties at once, we focus on a simple case, where the fine scale and the coarse-scale models can be written similarly, in the form of a simple elliptic partial differential equation in divergence form. The fine scale model includes heterogeneities at a small scale, a situation which is formalized by the fact that the coefficients in the fine scale model vary on a small length scale. After homogenization, this model yields an effective, macroscopic model, which includes no small scale. In many cases, a sound theoretical groundwork exists for such homogenization results. The difficulty stems from the fact that the models generally lead to prohibitively costly computations. For such a case, simple from the theoretical viewpoint, our aim is to focus on different practical computational approaches to speed-up the computations. One possibility, among others, is to look for specific random materials, relevant from the practical viewpoint, and for which a dedicated approach can be proposed, that is less expensive than the general approach.

MATHRISK Project-Team

3. Research Program

3.1. Risk management: modeling and optimization

3.1.1. Contagion modeling and systemic risk

After the recent financial crisis, systemic risk has emerged as one of the major research topics in mathematical finance. Interconnected systems are subject to contagion in time of distress. The scope is to understand and model how the bankruptcy of a bank (or a large company) may or not induce other bankruptcies. By contrast with the traditional approach in risk management, the focus is no longer on modeling the risks faced by a single financial institution, but on modeling the complex interrelations between financial institutions and the mechanisms of distress propagation among these.

The mathematical modeling of default contagion, by which an economic shock causing initial losses and default of a few institutions is amplified due to complex linkages, leading to large scale defaults, can be addressed by various techniques, such as network approaches (see in particular R. Cont et al. [45] and A. Minca [80]) or mean field interaction models (Garnier-Papanicolaou-Yang [73]).

We have contributed in the last years to the research on the control of contagion in financial systems in the framework of random graph models : In [46], [81], [5], A. Sulem with A. Minca and H. Amini consider a financial network described as a weighted directed graph, in which nodes represent financial institutions and edges the exposures between them. The distress propagation is modeled as an epidemics on this graph. They study the optimal intervention of a lender of last resort who seeks to make equity infusions in a banking system prone to insolvency and to bank runs, under complete and incomplete information of the failure cluster, in order to minimize the contagion effects. The paper [5] provides in particular important insight on the relation between the value of a financial system, connectivity and optimal intervention.

The results show that up to a certain connectivity, the value of the financial system increases with connectivity. However, this is no longer the case if connectivity becomes too large. The natural question remains how to create incentives for the banks to attain an optimal level of connectivity. This is studied in [58], where network formation for a large set of financial institutions represented as nodes is investigated. Linkages are source of income, and at the same time they bear the risk of contagion, which is endogeneous and depends on the strategies of all nodes in the system. The optimal connectivity of the nodes results from a game. Existence of an equilibrium in the system and stability properties is studied. The results suggest that financial stability is best described in terms of the mechanism of network formation than in terms of simple statistics of the network topology like the average connectivity.

3.1.2. Liquidity risk and Market Microstructure

Liquidity risk is the risk arising from the difficulty of selling (or buying) an asset. Usually, assets are quoted on a market with a Limit Order Book (LOB) that registers all the waiting limit buy and sell orders for this asset. The bid (resp. ask) price is the most expensive (resp. cheapest) waiting buy or sell order. If a trader wants to sell a single asset, he will sell it at the bid price, but if he wants to sell a large quantity of assets, he will have to sell them at a lower price in order to match further waiting buy orders. This creates an extra cost, and raises important issues. From a short-term perspective (from few minutes to some days), it may be interesting to split the selling order and to focus on finding optimal selling strategies. This requires to model the market microstructure, i.e. how the market reacts in a short time-scale to execution orders. From a long-term perspective (typically, one month or more), one has to understand how this cost modifies portfolio managing strategies (especially delta-hedging or optimal investment strategies). At this time-scale, there is no need to model precisely the market microstructure, but one has to specify how the liquidity costs aggregate.

For rather liquid assets, liquidity risk is usually taken into account via price impact models which describe how a (large) trader influences the asset prices. Then, one is typically interested in the optimal execution problem: how to buy/sell a given amount of assets optimally within a given deadline. This issue is directly related to the existence of statistical arbitrage or Price Manipulation Strategies (PMS). Most of price impact models deal with single assets. A. Alfonsi, F. Klöck and A. Schied [44] have proposed a multi-assets price impact model that extends previous works. Price impact models are usually relevant when trading at an intermediary frequency (say every hour). At a lower frequency, price impact is usually ignored while at a high frequency (every minute or second), one has to take into account the other traders and the price jumps, tick by tick. Midpoint price models are thus usually preferred at this time scale. With P. Blanc, Alfonsi [3] has proposed a model that makes a bridge between these two types of model: they have considered an (Obizhaeva and Wang) price impact model, in which the flow of market orders generated by the other traders is given by an exogeneous process. They have shown that Price Manipulation Strategies exist when the flow of order is a compound Poisson process. However, modeling this flow by a mutually exciting Hawkes process with a particular parametrization allows them to exclude these PMS. Besides, the optimal execution strategy is explicit in this model. A practical implementation is given in [40].

3.1.3. Dependence modeling

- Calibration of stochastic and local volatility models. The volatility is a key concept in modern mathematical finance, and an indicator of market stability. Risk management and associated instruments depend strongly on the volatility, and volatility modeling is a crucial issue in the finance industry. Of particular importance is the assets *dependence* modeling.

By Gyongy's theorem, a local and stochastic volatility model is calibrated to the market prices of all call options with positive maturities and strikes if its local volatility function is equal to the ratio of the Dupire local volatility function over the root conditional mean square of the stochastic volatility factor given the spot value. This leads to a SDE nonlinear in the sense of McKean. Particle methods based on a kernel approximation of the conditional expectation, as presented by Guyon and Henry-Labordère [74], provide an efficient calibration procedure even if some calibration errors may appear when the range of the stochastic volatility factor is very large. But so far, no existence result is available for the SDE nonlinear in the sense of McKean. In the particular case when the local volatility function is equal to the inverse of the root conditional mean square of the stochastic volatility factor multiplied by the spot value given this value and the interest rate is zero, the solution to the SDE is a fake Brownian motion. When the stochastic volatility factor is a constant (over time) random variable taking finitely many values and the range of its square is not too large, B. Jourdain and A. Zhou proved existence to the associated Fokker-Planck equation [21]. Thanks to results obtained by Figalli in [69], they deduced existence of a new class of fake Brownian motions. They extended these results to the special case of the LSV model called Regime Switching Local Volatility, when the stochastic volatility factor is a jump process taking finitely many values and with jump intensities depending on the spot level.

- Interest rates modeling. Affine term structure models have been popularized by Dai and Singleton [59], Duffie, Filipovic and Schachermayer [60]. They consider vector affine diffusions (the coordinates are usually called factors) and assume that the short interest rate is a linear combination of these factors. A model of this kind is the Linear Gaussian Model (LGM) that considers a vector Ornstein-Uhlenbeck diffusions for the factors, see El Karoui and Lacoste [68]. A. Alfonsi et al. [37] have proposed an extension of this model, when the instantaneous covariation between the factors is given by a Wishart process. Doing so, the model keeps its affine structure and tractability while generating smiles for option prices. A price expansion around the LGM is obtained for Caplet and Swaption prices.

3.1.4. Robust finance

- Numerical Methods for Martingale Optimal Transport problems.

The Martingale Optimal Transport (MOT) problem introduced in [57] has received a recent attention in finance since it gives model-free hedges and bounds on the prices of exotic options. The market prices of liquid call and put options give the marginal distributions of the underlying asset at each traded maturity. Under the simplifying assumption that the risk-free rate is zero, these probability measures are in increasing convex

order, since by Strassen's theorem this property is equivalent to the existence of a martingale measure with the right marginal distributions. For an exotic payoff function of the values of the underlying on the time-grid given by these maturities, the model-free upper-bound (resp. lower-bound) for the price consistent with these marginal distributions is given by the following martingale optimal transport problem : maximize (resp. minimize) the integral of the payoff with respect to the martingale measure over all martingale measures with the right marginal distributions. Super-hedging (resp. sub-hedging) strategies are obtained by solving the dual problem. With J. Corbetta, A. Alfonsi and B. Jourdain [12] have studied sampling methods preserving the convex order for two probability measures μ and ν on $\mathbf{R}^d$, with ν dominating μ.

Their method is the first generic approach to tackle the martingale optimal transport problem numerically and can also be applied to several marginals.

- Robust option pricing in financial markets with imperfections.

A. Sulem, M.C. Quenez and R. Dumitrescu have studied robust pricing in an imperfect financial market with default. The market imperfections are taken into account via the nonlinearity of the wealth dynamics. In this setting, the pricing system is expressed as a nonlinear g-expectation $\mathcal{E}^g$ induced by a nonlinear BSDE with nonlinear driver g and default jump (see [61]). A large class of imperfect market models can fit in this framework, including imperfections coming from different borrowing and lending interest rates, taxes on profits from risky investments, or from the trading impact of a large investor seller on the market prices and the default probability. Pricing and superhedging issues for American and game options in this context and their links with optimal stopping problems and Dynkin games with nonlinear expectation have been studied. These issues have also been addressed in the case of model uncertainty, in particular uncertainty on the default probability. The seller's robust price of a game option has been characterized as the value function of a Dynkin game under $\mathcal{E}^g$ expectation as well as the solution of a nonlinear doubly reflected BSDE in [9]. Existence of robust superhedging strategies has been studied. The buyer's point of view and arbitrage issues have also been studied in this context.

In a Markovian framework, the results of the paper [8] on combined optimal stopping/stochastic control with $\mathcal{E}^g$ expectation allows us to address American nonlinear option pricing when the payoff function is only Borelian and when there is ambiguity both on the drift and the volatility of the underlying asset price process. Robust optimal stopping of dynamic risk measures induced by BSDEs with jumps with model ambiguity is studied in [83].

3.2. Perspectives in Stochastic Analysis

3.2.1. Optimal transport and longtime behavior of Markov processes

The dissipation of general convex entropies for continuous time Markov processes can be described in terms of backward martingales with respect to the tail filtration. The relative entropy is the expected value of a backward submartingale. In the case of (non necessarily reversible) Markov diffusion processes, J. Fontbona and B. Jourdain [71] used Girsanov theory to explicit the Doob-Meyer decomposition of this submartingale. They deduced a stochastic analogue of the well known entropy dissipation formula, which is valid for general convex entropies, including the total variation distance. Under additional regularity assumptions, and using Itô's calculus and ideas of Arnold, Carlen and Ju [47], they obtained a new Bakry-Emery criterion which ensures exponential convergence of the entropy to 0. This criterion is non-intrinsic since it depends on the square root of the diffusion matrix, and cannot be written only in terms of the diffusion matrix itself. They provided examples where the classic Bakry Emery criterion fails, but their non-intrinsic criterion applies without modifying the law of the diffusion process.

With J. Corbetta, A. Alfonsi and B. Jourdain have studied the time derivative of the Wasserstein distance between the marginals of two Markov processes [11]. The Kantorovich duality leads to a natural candidate for this derivative. Up to the sign, it is the sum of the integrals with respect to each of the two marginals of the corresponding generator applied to the corresponding Kantorovich potential. For pure jump processes with bounded intensity of jumps, J. Corbetta, A. Alfonsi and B. Jourdain [41] proved that the evolution of the Wasserstein distance is actually given by this candidate. In dimension one, they showed that this remains

true for Piecewise Deterministic Markov Processes. They applied the formula to estimate the exponential decrease rate of the Wasserstein distance between the marginals of two birth and death processes with the same generator in terms of the Wasserstein curvature.

3.2.2. Mean-field systems: modeling and control

- Mean-field limits of systems of interacting particles. In [77], B. Jourdain and his former PhD student J. Reygner have studied a mean-field version of rank-based models of equity markets such as the Atlas model introduced by Fernholz in the framework of Stochastic Portfolio Theory. They obtained an asymptotic description of the market when the number of companies grows to infinity. Then, they discussed the long-term capital distribution, recovering the Pareto-like shape of capital distribution curves usually derived from empirical studies, and providing a new description of the phase transition phenomenon observed by Chatterjee and Pal. They have also studied multitype sticky particle systems which can be obtained as vanishing noise limits of multitype rank-based diffusions (see [76]). Under a uniform strict hyperbolicity assumption on the characteristic fields, they constructed a multitype version of the sticky particle dynamics. In [78], they obtain the optimal rate of convergence as the number of particles grows to infinity of the approximate solutions to the diagonal hyperbolic system based on multitype sticky particles and on easy to compute time discretizations of these dynamics.

In [72], N. Fournier and B. Jourdain are interested in the two-dimensional Keller-Segel partial differential equation. This equation is a model for chemotaxis (and for Newtonian gravitational interaction).

- Mean field control and Stochastic Differential Games (SDGs). To handle situations where controls are chosen by several agents who interact in various ways, one may use the theory of Stochastic Differential Games (SDGs). Forward–Backward SDG and stochastic control under Model Uncertainty are studied in [84] by A. Sulem and B. Øksendal. Also of interest are large population games, where each player interacts with the average effect of the others and individually has negligible effect on the overall population. Such an interaction pattern may be modeled by mean field coupling and this leads to the study of mean-field stochastic control and related SDGs. A. Sulem, Y. Hu and B. Øksendal have studied singular mean field control problems and singular mean field two-players stochastic differential games [75]. Both sufficient and necessary conditions for the optimal controls and for the Nash equilibrium are obtained. Under some assumptions, the optimality conditions for singular mean-field control are reduced to a reflected Skorohod problem. Applications to optimal irreversible investments under uncertainty have been investigated. Predictive mean-field equations as a model for prices influenced by beliefs about the future are studied in [86].

3.2.3. Stochastic control and optimal stopping (games) under nonlinear expectation

M.C. Quenez and A. Sulem have studied optimal stopping with nonlinear expectation $\mathcal{E}^g$ induced by a BSDE with jumps with nonlinear driver g and irregular obstacle/payoff (see [83]). In particular, they characterize the value function as the solution of a reflected BSDE. This property is used in [67] to address American option pricing in markets with imperfections. The Markovian case is treated in [64] when the payoff function is continuous.

In [8], M.C. Quenez, A. Sulem and R. Dumitrescu study a combined optimal control/stopping problem under nonlinear expectation $\mathcal{E}^g$ in a Markovian framework when the terminal reward function is only Borelian. In this case, the value function u associated with this problem is irregular in general. They establish a *weak* dynamic programming principle (DPP), from which they derive that the upper and lower semi-continuous envelopes of u are the sub- and super- *viscosity solution* of an associated nonlinear Hamilton-Jacobi-Bellman variational inequality.

The problem of a generalized Dynkin game problem with nonlinear expectation $\mathcal{E}^g$ is addressed in [65]. Under Mokobodzki's condition, we establish the existence of a value function for this game, and characterize this value as the solution of a doubly reflected BSDE. The results of this work are used in [9] to solve the problem of game option pricing in markets with imperfections.

A generalized mixed game problem when the players have two actions: continuous control and stopping is studied in a Markovian framework in [66]. In this work, dynamic programming principles (DPP) are established: a strong DPP is proved in the case of a regular obstacle and a weak one in the irregular case. Using these DPPs, links with parabolic partial integro-differential Hamilton-Jacobi- Bellman variational inequalities with two obstacles are obtained.

With B. Øksendal and C. Fontana, A. Sulem has contributed on the issues of robust utility maximization [85], [86], and relations between information and performance [70].

3.2.4. Generalized Malliavin calculus

Vlad Bally has extended the stochastic differential calculus built by P. Malliavin which allows one to obtain integration by parts and associated regularity probability laws. In collaboration with L. Caramellino (Tor Vegata University, Roma), V. Bally has developed an abstract version of Malliavin calculus based on a splitting method (see [49]). It concerns random variables with law locally lower bounded by the Lebesgue measure (the so-called Doeblin's condition). Such random variables may be represented as a sum of a "smooth" random variable plus a rest. Based on this smooth part, he achieves a stochastic calculus which is inspired from Malliavin calculus [6]. An interesting application of such a calculus is to prove convergence for irregular test functions (total variation distance and more generally, distribution distance) in some more or less classical frameworks as the Central Limit Theorem, local versions of the CLT and moreover, general stochastic polynomials [53]. An exciting application concerns the number of roots of trigonometric polynomials with random coefficients [15]. Using Kac Rice lemma in this framework one comes back to a multidimensional CLT and employs Edgeworth expansions of order three for irregular test functions in order to study the mean and the variance of the number of roots. Another application concerns U statistics associated to polynomial functions. The techniques of generalized Malliavin calculus developed in [49] are applied in for the approximation of Markov processes (see [56] and [55]). On the other hand, using the classical Malliavin calculus, V. Bally in collaboration with L. Caramellino and P. Pigato studied some subtle phenomena related to diffusion processes, as short time behavior and estimates of tubes probabilities (see [51], [52], [50]).

3.3. Numerical Probability

Our project team is very much involved in numerical probability, aiming at pushing numerical methods towards the effective implementation. This numerical orientation is supported by a mathematical expertise which permits a rigorous analysis of the algorithms and provides theoretical support for the study of rates of convergence and the introduction of new tools for the improvement of numerical methods. This activity in the MathRisk team is strongly related to the development of the Premia software.

3.3.1. Simulation of stochastic differential equations

3.3.1.1. - Weak convergence of the Euler scheme in optimal transport distances.

With A. Kohatsu-Higa, A. Alfonsi and B. Jourdain [4] have proved using optimal transport tools that the Wasserstein distance between the time marginals of an elliptic SDE and its Euler discretization with N steps is not larger than $C\sqrt{\log(N)}/N$. The logarithmic factor may is removed when the uniform time-grid is replaced by a grid still counting N points but refined near the origin of times.

3.3.1.2. - Strong convergence properties of the Ninomiya Victoir scheme and multilevel Monte-Carlo estimators.

With their former PhD student, A. Al Gerbi, E. Clément and B. Jourdain [1] have proved strong convergence with order $1/2$ of the Ninomiya-Victoir scheme which is known to exhibit order 2 of weak convergence [82]. This study was aimed at analysing the use of this scheme either at each level or only at the finest level of a multilevel Monte Carlo estimator : indeed, the variance of a multilevel Monte Carlo estimator is related to the strong error between the two schemes used in the coarse and fine grids at each level. In [38], they proved that the order of strong convergence of the crude Ninomiya Victoir scheme is improved to 1 when the vector fields corresponding to each Brownian coordinate in the SDE commute, and in [39], they studied the error introduced by discretizing the ordinary differential equations involved in the Ninomiya-Victoir scheme.

3.3.1.3. - Non-asymptotic error bounds for the multilevel Monte Carlo Euler method.

A. Kebaier and B. Jourdain are interested in deriving non-asymptotic error bounds for the multilevel Monte Carlo method. As a first step, they dealt in [20] with the explicit Euler discretization of stochastic differential equations with a constant diffusion coefficient. They obtained Gaussian-type concentration. To do so, they used the Clark-Ocone representation formula and derived bounds for the moment generating functions of the squared difference between a crude Euler scheme and a finer one and of the squared difference of their Malliavin derivatives. The estimation of such differences is much more complicated than the one of a single Euler scheme contribution and explains why they suppose the diffusion coefficient to be constant. This assumption ensures boundedness of the Malliavin derivatives of both the SDE and its Euler scheme.

3.3.1.4. - Computation of sensibilities of integrals with respect to the invariant measure.

In [48], R. Assaraf, B. Jourdain, T. Lelièvre and R. Roux considered the solution to a stochastic differential equation with constant diffusion coefficient and with a drift function which depends smoothly on some real parameter λ, and admitting a unique invariant measure for any value of λ around $\lambda = 0$. Their aim was to compute the derivative with respect to λ of averages with respect to the invariant measure, at $\lambda = 0$. They analyzed a numerical method which consists in simulating the process at $\lambda = 0$ together with its derivative with respect to λ on a long time horizon. They gave sufficient conditions implying uniform-in-time square integrability of this derivative. This allows in particular to compute efficiently the derivative with respect to λ of the mean of an observable through Monte Carlo simulations.

3.3.1.5. - Approximation of doubly reflected Backward stochastic differential equations.

R. Dumitrescu and C. Labart have studied the discrete time approximation scheme for the solution of a doubly reflected Backward Stochastic Differential Equation with jumps, driven by a Brownian motion and an independent compensated Poisson process [63], [62].

3.3.1.6. - Parametrix methods.

V. Bally and A. Kohatsu-Higa have recently proposed an unbiased estimator based on the parametrix method to compute expectations of functions of a given SDE ([54]). This method is very general, and A. Alfonsi, A. Kohastu-Higa and M. Hayashi [42] have applied it to the case of one-dimensional reflected diffusions. In this case, the estimator can be obtained explicitly by using the scheme of Lépingle [79] and is quite simple to implement. It is compared to other simulation methods for reflected SDEs.

3.3.2. Estimation of the parameters of a Wishart process

A. Alfonsi, A. Kebaier and C. Rey [43] have computed the Maximum Likelihood Estimator for the Wishart process and studied its convergence in the ergodic and in some non ergodic cases. In the ergodic case, which is the most relevant for applications, they obtain the standard square-root convergence. In the non ergodic case, the analysis rely on refined results for the Laplace transform of Wishart processes, which are of independent interest.

3.3.3. Optimal stopping and American options

In joint work with A. Bouselmi, D. Lamberton studied the asymptotic behavior of the exercise boundary near maturity for American put options in exponential Lévy models. In [7], they deal with jump-diffusion models, and establish that, in some cases, the behavior differs from the classical Black and Scholes setting. D. Lamberton has also worked on the binomial approximation of the American put. The conjectured rate of convergence is $O(1/n)$ where n is the number of time periods. He was able to derive a $O((\ln n)^{\alpha}/n)$ bound, where the exponent α is related to the asymptotic behavior of the exercise boundary near maturity.

MCTAO Project-Team

3. Research Program

3.1. Control Problems

McTAO's major field of expertise is control theory in the large sense. Let us give an overview of this field.

Modelling. Our effort is directed toward efficient methods for the control of real (physical) *systems*, based on a *model* of the system to be controlled. Choosing accurate models yet simple enough to allow control design is in itself a key issue. The typical continuous-time model is of the form $\mathrm{d}x/\mathrm{d}t = f(x,u)$ where x is the *state*, ideally finite dimensional, and u the *control*; the control is left free to be a function of time, or a function of the state, or obtained as the solution of another dynamical system that takes x as an input. Modelling amounts to deciding the nature and dimension of x, as well as the dynamics (roughly speaking the function f). Connected to modeling is identification of parameters when a finite number of parameters are left free in "f".

Controllability, path planning. Controllability is a property of a control system (in fact of a model) that two states in the state space can be connected by a trajectory generated by some control, here taken as an explicit function of time. Deciding on local or global controllability is still a difficult open question in general. In most cases, controllability can be decided by linear approximation, or non-controllability by "physical" first integrals that the control does not affect. For some critically actuated systems, it is still difficult to decide local or global controllability, and the general problem is anyway still open. Path planning is the problem of constructing the control that actually steers one state to another.

Optimal control. In optimal control, one wants to find, among the controls that satisfy some contraints at initial and final time (for instance given initial and final state as in path planning), the ones that minimize some criterion. This is important in many control engineering problems, because minimizing a cost is often very relevant. Mathematically speaking, optimal control is the modern branch of the calculus of variations, rather well established and mature [71], [47], [34], but with a lot of hard open questions. In the end, in order to actually compute these controls, ad-hoc numerical schemes have to be derived for effective computations of the optimal solutions. See more about our research program in optimal control in section 3.2 .

Feedback control. In the above two paragraphs, the control is an explicit function of time. To address in particular the stability issues (sensitivity to errors in the model or the initial conditions for example), the control has to be taken as a function of the (measured) state, or part of it. This is known as closed-loop control; it must be combined with optimal control in many real problems. On the problem of stabilization, there is longstanding research record from members of the team, in particular on the construction of "Control Lyapunov Functions", see [62], [73].

Classification of control systems One may perform various classes of transformations acting on systems, or rather on models... The simpler ones come from point-to-point transformations (changes of variables) on the state and control, and more intricate ones consist in embedding an extraneous dynamical system into the model, these are dynamic feedback transformations, they change the dimension of the state. In most problems, choosing the proper coordinates, or the right quantities that describe a phenomenon, sheds light on a path to the solution; these proper choices may sometimes be found from an understanding of the modelled phenomenons, or it can come from the study of the geometry of the equations and the transformation acting on them. This justifies the investigations of these transformations on models for themselves. These topics are central in control theory; they are present in the team, see for instance the classification aspect in [52] or —although this research has not been active very recently— the study [70] of dynamic feedback and the so-called "flatness" property [65].

3.2. Optimal Control and its Geometry

Let us detail our research program concerning optimal control. Relying on Hamiltonian dynamics is now prevalent, instead of the Lagrangian formalism in classical calculus of variations. The two points of view run parallel when computing geodesics and shortest path in Riemannian Geometry for instance, in that there is a clear one-to-one correspondance between the solutions of the geodesic equation in the tangent bundle and the solution of the Pontryagin Maximum Principle in the cotangent bundle. In most optimal control problems, on the contrary, due to the differential constraints (velocities of feasible trajectories do not cover all directions in the state space), the Lagrangian formalism becomes more involved, while the Pontryagin Maximum Principle keeps the same form, its solutions still live in the cotangent bundle, their projections are the extremals, and a minimizing curve must be the projection of such a solution.

Cut and conjugate loci. The cut locus —made of the points where the extremals lose optimality— is obviously crucial in optimal control, but usually out of reach (even in low dimensions), and anyway does not have an analytic characterization because it is a non-local object. Fortunately, conjugate points —where the extremals lose *local* optimality— can be effectively computed with high accuracy for many control systems. Elaborating on the seminal work of the Russian and French schools (see [76], [35], [36] and [53] among others), efficient algorithms were designed to treat the smooth case. This was the starting point of a series of papers of members of the team culminating in the outcome of the *cotcot* software [46], followed by the *Hampath* [55] code. Over the years, these codes have allowed for the computation of conjugate loci in a wealth of situations including applications to space mechanics, quantum control, and more recently swimming at low Reynolds number. With in mind the two-dimensional analytic Riemannian framework, a heuristic approach to the global issue of determining cut points is to search for singularities of the conjugate loci; this line is however very delicate to follow on problems stemming from applications in three or more dimensions (see *e.g.* [56] and [43]). In all these situations, the fundamental object underlying the analysis is the curvature tensor. In Hamiltonian terms, one considers the dynamics of subspaces (spanned by Jacobi fields) in the Lagrangian Grassmannian [33]. This point of view withstands generalizations far beyond the smooth case: In L^1-minimization, for instance, discontinuous curves in the Grassmannian have to be considered (instantaneous rotations of Lagrangian subspaces still obeying symplectic rules [60]). The cut locus is a central object in Riemannian geometry, control and optimal transport. This is the motivation for a series of conferences on "The cut locus: A bridge over differential geometry, optimal control, and transport", co-organized by team members and Japanese colleagues, the next one should take place in Nice in 2020.

Riemann and Finsler geometry. Studying the distance and minimising geodesics in Riemannian Geometry or Finsler Geometry is a particular case of optimal control, simpler because there are no differential constraints; it is studied in the team for the following two reasons. On the one hand, after some tranformations, like averaging (see section 3.2) or reduction, some more difficult optimal control problems lead to a Riemann or Finsler geometry problem. On the other hand, optimal control, mostly the Hamiltonian setting, brings a fresh viewpoint on problems in Riemann and Finsler geometry. On Riemannian ellipsoids of revolution, the optimal control approach allowed to decide on the convexity of the injectivity domain, which, associated with non-negativity of the Ma-Trudinger-Wang curvature tensor, ensures continuity of the optimal transport on the ambient Riemannian manifold [64], [63]. The analysis in the oblate geometry [44] was completed in [59] in the prolate one, including a preliminary analysis of non-focal domains associated with conjugate loci. Averaging in systems coming from space mechanics control (see sections 3.2 and 4.1) with L^2-minimization yields a Riemannian metric, thoroughly computed in [42] together with its geodesic flow; in reduced dimension, its conjugate and cut loci were computed in [45] with Japanese Riemannian geometers. Averaging the same systems for minimum time yields a Finsler Metric, as noted in [41]. In [51], the geodesic convexity properties of these two types of metrics were compared. When perturbations (other than the control) are considered, they introduce a "drift", *i.e.* the Finsler metric is no longer symmetric.

Sub-Riemannian Geometry. Optimal control problems that pertain to sub-Riemannian Geometry bear all the difficulties of optimal control, like the role of singular/abnormal trajectories, while having some useful structure. They lead to many open problems, like smoothness of minimisers, see the recent monograph [69] for an introduction. Let us detail one open question related to these singular trajectories: the Sard conjecture in sub-Riemannian geometry. Given a totally non-holonomic distribution on a smooth manifold, the Sard Conjecture is concerned with the size of the set of points that can be reached by singular horizontal paths starting from a given point. In the setting of rank-two distributions in dimension three, the Sard conjecture is that this set should be a subset of the so-called Martinet surface, indeed small both in measure and in dimension. In [39], it has been proved that the conjecture holds in the case where the Martinet surface is smooth. Moreover, the case of singular real-analytic Martinet surfaces was also addressed. In this case, it was shown that the Sard Conjecture holds true under an assumption of non-transversality of the distribution on the singular set of the Martinet surface. It is, of course, very intersting to get rid of the remaining technical assumption, or to go to higher dimension. Note that any Sard-type result has strong consequences on the regularity of sub-Riemannian distance functions and in turn on optimal transport problems in the sub-Riemannian setting.

Small controls and conservative systems, averaging. Using averaging techniques to study small perturbations of integrable Hamiltonian systems is as old an idea as celestial mechanics. It is very subtle in the case of multiple periods but more elementary in the single period case, here it boils down to taking the average of the perturbation along each periodic orbit [37], [75]. This line of research stemmed out of applications to space engineering (see section 4.1): the control of the super-integrable Keplerian motion of a spacecraft orbiting around the Earth is an example of a slow-fast controlled system. Since weak propulsion is used, the control itself acts as a perturbation, among other perturbations of similar magnitudes: higher order terms of the Earth potential (including J_2 effect, first), potential of more distant celestial bodies (such as the Sun and the Moon), atmospheric drag, or even radiation pressure. Properly qualifying the convergence properties (when the small parameter goes to zero) is important and is made difficult by the presence of control. In [41], convergence is seen as convergence to a differential inclusion; this applies to minimum time; a contribution of this work is to put forward the metric character of the averaged system by yielding a Finsler metric (see section 3.2). Proving convergence of the extremals (solutions of the Pontryagin Maximum Principle) is more intricate. In [58], standard averaging ([37], [75]) is performed on the minimum time extremal flow after carefully identifying slow variables of the system thanks to a symplectic reduction. This alternative approach allows to retrieve the previous metric approximation, and to partly address the question of convergence. Under suitable assumptions on a given geodesic of the averaged system (disconjugacy conditions, namely), one proves existence of a family of quasi-extremals for the original system that converge towards the geodesic when the small perturbation parameter goes to zero. This needs to be improved, but convergence of all extremals to extremals of an "averaged Pontryagin Maximum Principle" certainly fails. In particular, one cannot hope for C^1-regularity on the value function when the small parameter goes to zero as swallowtail-like singularities due to the structure of local minima in the problem are expected. (A preliminary analysis has been made in [57].)

Optimality of periodic solutions/periodic controls. When seeking to minimize a cost with the constraint that the controls and/or part of the states are periodic (and with other initial and final conditions), the notion of conjugate points is more difficult than with straightforward fixed initial point. In [48], for the problem of optimizing the efficiency of the displacement of some micro-swimmers (see section 4.3) with periodic deformations, we used the sufficient optimality conditions established by R. Vinter's group [80], [66] for systems with non unique minimizers due to the existence of a group of symmetry (always present with a periodic minimizer-candidate control). This takes place in a long term collaboration with P. Bettiol (Univ. Bretagne Ouest) on second order sufficient optimality conditions for periodic solutions, or in the presence of higher dimensional symmetry groups, following [80], [66]. Another question relevant to locomotion is the following. Observing animals (or humans), or numerically solving the optimal control problem associated with driftless micro-swimmers for various initial and final conditions, we remark that the optimal strategies of deformation seem to be periodic, at least asymptotically for large distances. This observation is the starting point for characterizing dynamics for which some optimal solutions are periodic, and asymptotically attract

other solutions as the final time grows large; this is reminiscent of the "turnpike theorem" (classical, recently applied to nonlinear situations in [79]).

Software. These applications (but also the development of theory where numerical experiments can be very enlightening) require many algorithmic and numerical developments that are an important side of the team activity. The software *Hampath* (see section 6.1) is maintained by former members of the team in close collaboration with McTAO. We also use direct discretization approaches (such as the Bocop solver developed by COMMANDS) in parallel. Apart from this, we develop on-demand algorithms and pieces of software, for instance we have to interact with a production software developed by Thales Alenia Space. A strong asset of the team is the interplay of its expertise in geometric control theory with applications and algorithms (see sections 4.1 to 4.3) on one hand, and with optimal transport, and more recently Hamiltonian dynamics, on the other. In 2019, the ADT `ct` (Control Toolbox) has started with a first sprint in "AMDT mode" with Sophia SED during spring 2019. In addition to McTAO, researchers from the CAGE team (Inria Paris) and the APO team (CNRS Toulouse) are involved. The idea is to put together the efforts on BOCOP and HamPath to go towards a reference toolbox in optimal control. After the first sprint cycle (24 months being planned on the whole action), some starting points have been addressed including: continuous integration for BOCOP and HamPath, refresh on collaborative development tools, first steps of software refactoring, first test of a high-end interface (throught scripting, notebooks, or an *ad hoc* GUI). The next sprint is planned during spring 2020.

3.3. Optimal Transport

Given two measures, and calling transport maps the maps that transport the first measure into the second one, the Monge-Kantorovich problem of Optimal Transport is the search of the minimum of some cost on the set of transport maps. The cost of a map usually comes from some point to point cost and the transport measure. This topic attracted renewed attention in the last decade, and has ongoing applications of many types. Matching optimal transport with geometric control theory is one originality of our team. Let us sketch an important class of open problems. In collaboration with R. McCann [68], we worked towards identifying the costs that admit unique optimizers in the Monge-Kantorovich problem of optimal transport between arbitrary probability densities. For smooth costs and densities on compact manifolds, the only known examples for which the optimal solution is always unique require at least one of the two underlying spaces to be homeomorphic to a sphere. We have introduced a multivalued dynamics induced by the transportation cost between the target and source space, for which the presence or absence of a sufficiently large set of periodic trajectories plays a role in determining whether or not optimal transport is necessarily unique. This insight allows us to construct smooth costs on a pair of compact manifolds with arbitrary topology, so that the optimal transport between any pair of probability densities is unique. We investigated further this problem of uniquely minimizing costs and obtained in collaboration with Abbas Moameni [10] a result of density of uniquely minimizing costs in the C^0-topology. The results in higher topology should be the subject of some further research.

MEMPHIS Project-Team

3. Research Program

3.1. Reduced-order models

Massive parallelization and rethinking of numerical schemes will allow the use of mathematical models for a broader class of physical problems. For industrial applications, there is an increasing need for rapid and reliable numerical simulators to tackle design and control tasks. To provide a concrete example, in the design process of an aircraft, the flight conditions and manoeuvres, which provide the largest aircraft loads, are not known *a priori*. Therefore, the aerodynamic and inertial forces are calculated for a large number of conditions to give an estimate of the maximum loads, and hence stresses, that the structure of the detailed aircraft design might experience in service. As a result, the number of simulations required for a realistic design problem could easily be in the order of tens of millions. Even with simplistic models of the aircraft behavior this is an unfeasible number of separate simulations. However, engineering experience is used to identify the most likely critical load conditions, meaning that approximately hundreds of thousands simulations are required for conventional aircraft configurations. Furthermore, these analyses have to be repeated every time that there is an update in the aircraft structure.

Compared to existing approaches for ROMs [35], our interest will be focused on two axes. On the one hand, we start from the consideration that small, highly nonlinear scales are typically concentrated in limited spatial regions of the full simulation domain. So for example, in the flow past a wing, the highly non-linear phenomena take place in the proximity of the walls at the scale of a millimeter, for computational domains that are of the order of hundreds of meters. Based on these considerations, we propose in [31] a multi-scale model where the large scales are described by far-field models based on ROMs and the small scales are simulated by high-fidelity models. The whole point for this approach is to optimally decouple the far field from the near field.

A second characterizing feature of our ROM approach is non-linear interpolation. We start from the consideration that dynamical models derived from the projection of the PDE model in the reduced space are neither stable to numerical integration nor robust to parameter variation when hard non-linear multi-scale phenomena are considered.

However, thanks to Proper Orthogonal Decomposition (POD) [41], [47], [30] we can accurately approximate large solution databases using a low-dimensional base. Recent techniques to investigate the temporal evolution of the POD modes (Koopman modes [42], [28], Dynamic Mode Decomposition [45]) and allow a dynamic discrimination of the role played by each of them. This in turn can be exploited to interpolate between modes in parameter space, thanks to ideas relying on optimal transportation [50], [32] that we have started developing in the FP7 project FFAST and H2020 AEROGUST.

3.2. Hierarchical Cartesian schemes

We intend to conceive schemes that will simplify the numerical approximation of problems involving complex unsteady objects together with multi-scale physical phenomena. Rather than using extremely optimized but non-scalable algorithms, we adopt robust alternatives that bypass the difficulties linked to grid generation. Even if the mesh problem can be tackled today thanks to powerful mesh generators, it still represents a severe difficulty, in particular when highly complex unsteady geometries need to be dealt with. Industrial experience and common practice shows that mesh generation accounts for about 20% of overall analysis time, whereas creation of a simulation-specific geometry requires about 60%, and only 20% of overall time is actually devoted to analysis. The methods that we develop bypass the generation of tedious geometrical models by automatic implicit geometry representation and hierarchical Cartesian schemes.

The approach that we plan to develop combines accurate enforcement of unfitted boundary conditions with adaptive octree and overset grids. The core idea is to use an octree/overset mesh for the approximation of the solution fields, while the geometry is captured by level set functions [46], [40] and boundary conditions are imposed using appropriate interpolation methods [27], [49], [44]. This eliminates the need for boundary-conforming meshes that require time-consuming and error-prone mesh generation procedures, and opens the door for simulation of very complex geometries. In particular, it will be possible to easily import the industrial geometry and to build the associated level set function used for simulation.

Hierarchical octree grids offer several considerable advantages over classical adaptive mesh refinement for body-fitted meshes, in terms of data management, memory footprint and parallel HPC performance. Typically, when refining unstructured grids, like for example tetrahedral grids, it is necessary to store the whole data tree corresponding to successive subdivisions of the elements and eventually recompute the full connectivity graph. In the linear octree case that we develop, only the tree leaves are stored in a linear array, with a considerable memory advantage. The mapping between the tree leaves and the linear array as well as the connectivity graph is efficiently computed thanks to an appropriate space-filling curve. Concerning parallelization, linear octrees guarantee a natural load balancing thanks to the linear data structure, whereas classical unstructured meshes require sophisticated (and moreover time consuming) tools to achieve proper load distribution (SCOTCH, METIS etc.). Of course, using unfitted hierarchical meshes requires further development and analysis of methods to handle the refinement at level jumps in a consistent and conservative way, accuracy analysis for new finite-volume or finite-difference schemes, efficient reconstructions at the boundaries to recover appropriate accuracy and robustness. These subjects, that are currently virtually absent at Inria, are among the main scientific challenges of our team.

MEPHYSTO Team

3. Research Program

3.1. Time asymptotics: Stationary states, solitons, and stability issues

The team investigates the existence of solitons and their link with the global dynamical behavior for nonlocal problems such as that of the Gross–Pitaevskii (GP) equation which arises in models of dipolar gases. These models, in general, also introduce nonzero boundary conditions which constitute an additional theoretical and numerical challenge. Numerous results are proved for local problems, and numerical simulations allow to verify and illustrate them, as well as making a link with physics. However, most fundamental questions are still open at the moment for nonlocal problems.

The nonlinear Schrödinger (NLS) equation finds applications in numerous fields of physics. We concentrate, in a continued collaboration with our colleagues from the physics department (PhLAM) of the Université de Lille (UdL), in the framework of the Laboratoire d'Excellence CEMPI, on its applications in nonlinear optics and cold atom physics. Issues of orbital stability and modulational instability are central here.

Another typical example of problems that the team wishes to address concerns the Landau–Lifshitz (LL) equation, which describes the dynamics of the spin in ferromagnetic materials. This equation is a fundamental model in the magnetic recording industry [37] and solitons in magnetic media are of particular interest as a mechanism for data storage or information transfer [38]. It is a quasilinear PDE involving a function that takes values on the unit sphere $\mathbb{S}^2$ of $\mathbb{R}^3$. Using the stereographic projection, it can be seen as a quasilinear Schrödinger equation and the questions about the solitons, their dynamics and potential blow-up of solutions evoked above are also relevant in this context. This equation is less understood than the NLS equation: even the Cauchy theory is not completely done [36], [35]. In particular, the geometry of the target sphere imposes nonvanishing boundary conditions; even in dimension one, there are kink-type solitons having different limits at $\pm\infty$.

3.2. Derivation of macroscopic laws from microscopic dynamics

The team investigates, from a microscopic viewpoint, the dynamical mechanism at play in the phenomenon of relaxation towards thermal equilibrium for large systems of interacting particles. For instance, a first step consists in giving a rigorous proof of the fact that a particle repeatedly scattered by random obstacles through a Hamiltonian scattering process will eventually reach thermal equilibrium, thereby completing previous work in this direction by the team. As a second step, similar models as the ones considered classically will be defined and analysed in the quantum mechanical setting, and more particularly in the setting of quantum optics.

Another challenging problem is to understand the interaction of large systems with the boundaries, which is responsible for most energy exchanges (forcing and dissipation), even though it is concentrated in very thin layers. The presence of boundary conditions to evolution equations sometimes lacks understanding from a physical and mathematical point of view. In order to legitimate the choice done at the macroscopic level of the mathematical definition of the boundary conditions, we investigate systems of atoms (precisely chains of oscillators) with different local microscopic defects. We apply our recent techniques to understand how anomalous (in particular fractional) diffusive systems interact with the boundaries. For instance, the powerful tool given by Wigner functions that we already used has been successfully applied to the derivation of anomalous behaviors in open systems (for instance in [7]). The next step consists in developing an extension of that tool to deal with bounded systems provided with fixed boundaries. We also intend to derive anomalous diffusion by adding long range interactions to diffusive models. There are very few rigorous results in this direction. Finally, we aim at obtaining from a microscopic description the fractional porous medium equation (FPME), a nonlinear variation of the fractional diffusion equation, involving the fractional Laplacian instead of the usual one. Its rigorous study carries out many mathematical difficulties in treating at the same time the

nonlinearity and fractional diffusion. We want to make PDE theorists and probabilists work together, in order to take advantage of the analytical results which went far ahead and are more advanced than the statistical physics theory.

3.3. Numerical methods: analysis and simulations

The team addresses both questions of precision and numerical cost of the schemes for the numerical integration of nonlinear evolution PDEs, such as the NLS equation. In particular, we aim at developing, studying and implementing numerical schemes with high order that are more efficient for these problems. We also want to contribute to the design and analysis of schemes with appropriate qualitative properties. These properties may as well be "asymptotic preserving" properties, energy-preserving properties, or convergence to an equilibrium properties. Other numerical goals of the team include the numerical simulation of standing waves of nonlinear nonlocal GP equations. We also keep on developing numerical methods to efficiently simulate and illustrate theoretical results on instability, in particular in the context of the modulational instability in optical fibers, where we study the influence of randomness in the physical parameters of the fibers.

The team also designs simulation methods to estimate the accuracy of the physical description via microscopic systems, by computing precisely the rate of convergence as the system size goes to infinity. One method under investigation is related to cloning algorithms, which were introduced very recently and turn out to be essential in molecular simulation.

MINGUS Project-Team

3. Research Program

3.1. Research Program

The MINGUS project is devoted to the mathematical and numerical analysis of models arising in plasma physics and nanotechnology. The main goal is to construct and analyze numerical methods for the approximation of PDEs containing multiscale phenomena. Specific multiscale numerical schemes will be proposed and analyzed in different regimes (namely highly-oscillatory and dissipative). The ultimate goal is to dissociate the physical parameters (generically denoted by ε) from the numerical parameters (generically denoted by h) with a uniform accuracy. Such a task requires mathematical prerequisite of the PDEs.

Then, for a given stiff (highly-oscillatory or dissipative) PDE, the methodology of the MINGUS team will be the following

- Mathematical study of the asymptotic behavior of multiscale models.
  This part involves averaging and asymptotic analysis theory to derive asymptotic models, but also long-time behavior of the considered models.
- Construction and analysis of multiscale numerical schemes.
  This part is the core of the project and will be deeply inspired for the mathematical prerequisite. In particular, our ultimate goal is the design of *Uniformly Accurate* (UA) schemes, whose accuracy is independent of ε.
- Validation on physically relevant problems.
  The last goal of the MINGUS project is to validate the new numerical methods, not only on toy problems, but also on realistic models arising in physics of plasmas and nanotechnologies. We will benefit from the Selalib software library which will help us to scale-up our new numerical methods to complex physics.

3.1.1. Dissipative problems

In the dissipative context, the asymptotic analysis is quite well understood in the deterministic case and multiscale numerical methods have been developed in the last decades. Indeed, the so-called Asymptotic-Preserving schemes has retained a lot of attention all over the world, in particular in the context of collisional kinetic equations. But, there is still a lot of work to do if one is interesting in the derivation high order asymptotic models, which enable to capture the original solution for all time. Moreover, this analysis is still misunderstood when more complex systems are considered, involving non homogeneous relaxation rates or stochastic terms for instance. Following the methodology we aim at using, we first address the mathematical analysis before deriving multiscale efficient numerical methods.

A simple model of dissipative systems is governed by the following differential equations

$$\left\{ \begin{array}{rcll} \dfrac{dx^\varepsilon(t)}{dt} & = & \mathcal{G}(x^\varepsilon(t), y^\varepsilon(t)), & x^\varepsilon(0) = x_0, \\ \dfrac{dy^\varepsilon(t)}{dt} & = & -\dfrac{y^\varepsilon(t)}{\varepsilon} + \mathcal{H}(x^\varepsilon(t), y^\varepsilon(t)), & y^\varepsilon(0) = y_0, \end{array} \right. \tag{45}$$

for given initial condition $(x_0, y_0) \in \mathbb{R}^2$ and given smooth functions $\mathcal{G}, \mathcal{H}$ which possibly involve stochastic terms.

3.1.1.1. Asymptotic analysis of dissipative PDEs (F. Castella, P. Chartier, A. Debussche, E. Faou, M. Lemou)

Derivation of asymptotic problems

Our main goal is to analyze the asymptotic behavior of dissipative systems of the form (3) when ε goes to zero. The *center manifold theorem* [35] is of great interest but is largely unsatisfactory from the following points of view

- a constructive approach of h and x_0^ε is clearly important to identify the high-order asymptotic models: this would require expansions of the solution by means of B-series or word-series [37] allowing the derivation of error estimates between the original solution and the asymptotic one.
- a better approximation of the transient phase is strongly required to capture the solution for small time: extending the tools developed in averaging theory, the main goal is to construct a suitable change of variable which enables to approximate the original solution for all time.

Obviously, even at the ODE level, a deep mathematical analysis has to be performed to understand the asymptotic behavior of the solution of (3). But, the same questions arise at the PDE level. Indeed, one certainly expects that dissipative terms occurring in collisional kinetic equations (2) may be treated theoretically along this perspective. The key new point indeed is to see the center manifold theorem as a change of variable in the space on unknowns, while the standard point of view leads to considering the center manifold as an asymptotic object.

Stochastic PDEs
We aim at analyzing the asymptotic behavior of stochastic collisional kinetic problems, that is equation of the type (2). The noise can describe creation or absorption (as in (2)), but it may also be a forcing term or a random magnetic field. In the parabolic scaling, one expect to obtain parabolic SPDEs at the limit. More precisely, we want to understand the fluid limits of kinetic equations in the presence of noise. The noise is smooth and non delta correlated. It contains also a small parameter and after rescaling converges formally to white noise. Thus, this adds another scale in the multiscale analysis. Following the pioneering work [38], some substantial progresses have been done in this topic.

More realistic problems may be addressed such as high field limit describing sprays, or even hydrodynamic limit. The full Boltzmann equation is a very long term project and we wish to address simpler problems such as convergences of BGK models to a stochastic Stokes equation.

The main difficulty is that when the noise acts as a forcing term, which is a physically relevant situation, the equilibria are affected by the noise and we face difficulties similar to that of high field limit problems. Also, a good theory of averaging lemma in the presence of noise is lacking. The methods we use are generalization of the perturbed test function method to the infinite dimensional setting. We work at the level of the generator of the infinite dimensional process and prove convergence in the sense of the martingale problems. A further step is to analyse the speed of convergence. This is a prerequisite if one wants to design efficient schemes. This requires more refined tools and a good understanding of the Kolmogorov equation.

3.1.1.2. Numerical schemes for dissipative problems (All members)

The design of numerical schemes able to reproduce the transition from the microscopic to macroscopic scales largely matured with the emergence of the Asymptotic Preserving schemes which have been developed initially for collisional kinetic equations (actually, for solving (2) when $\beta \to 0$). Several techniques have flourished in the last decades. As said before, AP schemes entail limitations which we aim at overcoming by deriving

- AP numerical schemes whose numerical cost diminishes as $\beta \to 0$,
- Uniformly accurate numerical schemes, whose accuracy is independent of β.

Time diminishing methods
The main goal consists in merging Monte-Carlo techniques [33] with AP methods for handling *automatically* multiscale phenomena. As a result, we expect that the cost of the so-obtained method decreases when the asymptotic regime is approached; indeed, in the collisional (i.e. dissipative) regime, the deviational part becomes negligible so that a very few number of particles will be generated to sample it. A work in this direction has been done by members of the team.

We propose to build up a method which permits to realize the transition from the microscopic to the macroscopic description without domain decomposition strategies which normally oblige to fix and tune an interface in the physical space and some threshold parameters. Since it will permit to go over domain decomposition and AP techniques, this approach is a very promising research direction in the numerical approximation of multiscale kinetic problems arising in physics and engineering.

Uniformly accurate methods
To overcome the accuracy reduction observed in AP schemes for intermediate regimes, we intend to construct and analyse multiscale numerical schemes for (3) whose error is uniform with respect to ε. The construction of such a scheme requires a deep mathematical analysis as described above. Ideally one would like to develop schemes that preserve the center manifold (without computing the latter!) as well as schemes that resolve numerically the stiffness induced by the fast convergence to equilibrium (the so-called transient phase). First, our goal is to extend the strategy inspired by the central manifold theorem in the ODE case to the PDE context, in particular for collisional kinetic equations (2) when $\beta \to 0$. The design of Uniformly Accurate numerical schemes in this context would require to generalize two-scale techniques introduced in the framework of highly-oscillatory problems [36].

Multiscale numerical methods for stochastic PDEs
AP schemes have been developed recently for kinetic equations with noise in the context of Uncertainty Quantification UQ [41]. These two aspects (multiscale and UQ) are two domains which usually come within the competency of separate communities. UQ has drawn a lot of attention recently to control the propagation of data pollution; undoubtedly UQ has a lot of applications and one of our goals will be to study how sources of uncertainty are amplified or not by the multiscale character of the model. We also wish to go much further and by developing AP schemes when the noise is also rescaled and the limit is a white noise driven SPDE, as described in section (3.1.1.1). For simple nonlinear problem, this should not present much difficulties but new ideas will definitely be necessary for more complicated problems when noise deeply changes the asymptotic equation.

3.1.2. Highly-oscillatory problems

As a generic model for highly-oscillatory systems, we will consider the equation

$$\frac{du^\varepsilon(t)}{dt} = \mathcal{F}(t/\varepsilon, u^\varepsilon(t)), \quad u^\varepsilon(0) = u_0, \tag{46}$$

for a given u_0 and a given periodic function $\mathcal{F}$ (of period P w.r.t. its first variable) which possibly involves stochastic terms. Solution u^ε exhibits high-oscillations in time superimposed to a slow dynamics. Asymptotic techniques -resorting in the present context to *averaging* theory [45]- allow to decompose

$$u^\varepsilon(t) = \Phi_{t/\varepsilon} \circ \Psi_t \circ \Phi_0^{-1}(u_0), \tag{47}$$

into a fast solution component, the εP-periodic change of variable $\Phi_{t/\varepsilon}$, and a slow component, the flow Ψ_t of a non-stiff *averaged* differential equation. Although equation (5) can be satisfied only up to a small remainder, various methods have been recently introduced in situations where (4) is posed in $\mathbb{R}^n$ or for the Schrödinger equation (1).

In the asymptotic behavior $\varepsilon \to 0$, it can be advantageous to replace the original singularly perturbed model (for instance (1) or (2)) by an approximate model which does not contain stiffness any longer. Such reduced models can be derived using asymptotic analysis, namely averaging methods in the case of highly-oscillatory problems. In this project, we also plan to go beyond the mere derivation of limit models, by searching for better approximations of the original problem. This step is of mathematical interest *per se* but it also paves the way of the construction of multiscale numerical methods.

3.1.2.1. Asymptotic analysis of highly-oscillatory PDEs (All members)

Derivation of asymptotic problems

We intend to study the asymptotic behavior of highly-oscillatory evolution equations of the form (4) posed in an infinite dimensional Banach space.

Recently, the stroboscopic averaging has been extended to the PDE context, considering nonlinear Schrödinger equation (1) in the highly-oscillatory regime. A very exciting way would be to use this averaging strategy for highly-oscillatory kinetic problem (2) as those encountered in strongly magnetized plasmas. This turns out to be a very promising way to re-derive gyrokinetic models which are the basis of tokamak simulations in the physicists community. In contract with models derived in the literature (see [34]) which only capture the average with respect to the oscillations, this strategy allows for the complete recovery of the exact solution from the asymptotic (non stiff) model. This can be done by solving companion transport equation that stems naturally from the decomposition (5).

Long-time behavior of Hamiltonian systems
The study of long-time behavior of nonlinear Hamiltonian systems have received a lot of interest during the last decades. It enables to put in light some characteristic phenomena in complex situations, which are beyond the reach of numerical simulations. This kind of analysis is of great interest since it can provide very precise properties of the solution. In particular, we will focus on the dynamics of nonlinear PDEs when the initial condition is close to a stationary solution. Then, the long-time behavior of the solution is studied through mainly three axis

- *linear stability*: considering the linearized PDE, do we have stability of a stationary solution ? Do we have linear Landau damping around stable non homogeneous stationary states?
- *nonlinear stability*: under a criteria, do we have stability of a stationary solution in energy norm like in [42], and does this stability persist under numerical discretization? For example one of our goals is to address the question of the existence and stability of discrete travelling wave in space and time.
- do we have existence of damped solutions for the full nonlinear problem ? Around homogeneous stationary states, solutions scatter towards a modified stationary state (see [43], [39]). The question of existence of Landau damping effects around non homogeneous states is still open and is one of our main goal in the next future.

Asymptotic behavior of stochastic PDEs
The study of SPDEs has known a growing interest recently, in particular with the fields medal of M. Hairer in 2014. In many applications such as radiative transfer, molecular dynamics or simulation of optical fibers, part of the physical interactions are naturally modeled by adding supplementary random terms (the noise) to the initial deterministic equations. From the mathematical point of view, such terms change drastically the behavior of the system.

- In the presence of noise, highly-oscillatory dispersive equations presents new problems. In particular, to study stochastic averaging of the solution, the analysis of the long time behavior of stochastic dispersive equations is required, which is known to be a difficult problem in the general case. In some cases (for instance highly-oscillatory Schrödinger equation (1) with a time white noise in the regime $\varepsilon \ll 1$), it is however possible to perform the analysis and to obtain averaged stochastic equations. We plan to go further by considering more difficult problems, such as the convergence of a stochastic Klein-Gordon-Zakharov system to as stochastic nonlinear Schrödinger equation.
- The long-time behavior of stochastic Schrödinger equations is of great interest to analyze mathematically the validity of the Zakharov theory for wave turbulence (see [44]). The problem of wave turbulence can be viewed either as a deterministic Hamiltonian PDE with random initial data or a randomly forced PDEs where the stochastic forcing is concentrated in some part of the spectrum (in this sense it is expected to be a hypoelliptic problem). One of our goals is to test the validity the Zakharov equation, or at least to make rigorous the spectrum distribution spreading observed in the numerical experiments.

3.1.2.2. Numerical schemes for highly-oscillatory problems (All members)

This section proposes to explore numerical issues raised by highly-oscillatory nonlinear PDEs for which (4) is a prototype. Simulating a highly-oscillatory phenomenon usually requires to adapt the numerical parameters in order to solve the period of size ε so as to accurately simulate the solution over each period, resulting in a unacceptable execution cost. Then, it is highly desirable to derive numerical schemes able to advance the solution by a time step independent of ε. To do so, our goal is to construct *Uniformly Accurate* (UA) numerical schemes, for which the numerical error can be estimated by Ch^p (h being any numerical parameters) with C independent of ε and p the order of the numerical scheme.

Recently, such numerical methods have been proposed by members of the team in the highly-oscillatory context [36]. They are mainly based on a separation of the fast and slow variables, as suggested by the decomposition (5). An additional ingredient to prove the uniformly accuracy of the method for (4) relies on the search for an appropriate initial data which enables to make the problem smooth with respect to ε.

Such an approach is assuredly powerful since it provides a numerical method which enables to capture the high oscillations in time of the solution (and not only its average) even with a large time step. Moreover, in the asymptotic regime, the potential gain is of order $1/\varepsilon$ in comparison with standard methods, and finally averaged models are not able to capture the intermediate regime since they miss important information of the original problem. We are strongly convinced that this strategy should be further studied and extended to cope with some other problems. The ultimate goal being to construct a scheme for the original equation which degenerates automatically into a consistent approximation of the averaged model, without resolving it, the latter can be very difficult to solve.

- Space oscillations:
  When rapidly oscillating coefficients in **space** (*i.e.* terms of the form $a(x, x/\varepsilon)$) occur in elliptic or parabolic equations, homogenization theory and numerical homogenization are usually employed to handle the stiffness. However, these strategies are in general not accurate for all $\varepsilon \in]0, 1]$. Then, the construction of numerical schemes which are able to handle both regimes in an uniform way is of great interest. Separating fast and slow *spatial* scales merits to be explored in this context. The delicate issue is then to extend the choice suitable initial condition to an *appropriate choice of boundary conditions* of the augmented problem.
- Space-time oscillations:
  For more complex problems however, the recent proposed approaches fail since the main oscillations cannot be identified explicitly. This is the case for instance when the magnetic field B depends on t or x in (2) but also for many other physical problems. We then have to deal with the delicate issue of space-time oscillations, which is known to be a very difficult problem from a mathematical and a numerical point of view. To take into account the space-time mixing, a periodic motion has to be detected together with a phase S which possibly depends on the time and space variables. These techniques originate from **geometric optics** which is a very popular technique to handle highly-frequency waves.
- Geometrical properties:
  The questions related to the geometric aspects of multiscale numerical schemes are of crucial importance, in particular when long-time simulations are addressed (see [40]). Indeed, one of the main questions of geometric integration is whether intrinsic properties of the solution may be passed onto its numerical approximation. For instance, if the model under study is Hamiltonian, then the exact flow is symplectic, which motivates the design of symplectic numerical approximation. For practical simulations of Hamiltonian systems, symplectic methods are known to possess very nice properties (see [40]). It is important to combine multiscale techniques to geometric numerical integration. All the problems and equations we intend to deal with will be addressed with a view to preserve intrinsic geometric properties of the exact solutions and/or to approach the asymptotic limit of the system in presence of a small parameter. An example of a numerical method developed by members of the team is the multi-revolution method.
- Quasi-periodic case:

So far, numerical methods have been proposed for the periodic case with single frequency. However, the quasi-periodic case [0] is still misunderstood although many complex problems involve multi-frequencies. Even if the quasi-periodic averaging is doable from a theoretical point of view in the ODE case (see [45]), it is unclear how it can be extended to PDEs. One of the main obstacle being the requirement, usual for ODEs like (4), for $\mathcal{F}$ to be analytic in the periodic variables, an assumption which is clearly impossible to meet in the PDE setting. An even more challenging problem is then the design of numerical methods for this problem.

- extension to stochastic PDEs:
  All these questions will be revisited within the stochastic context. The mathematical study opens the way to the derivation of efficient multiscale numerical schemes for this kind of problems. We believe that the theory is now sufficiently well understood to address the derivation and numerical analysis of multiscale numerical schemes. Multi-revolution composition methods have been recently extended to highly-oscillatory stochastic differential equations The generalization of such multiscale numerical methods to SPDEs is of great interest. The analysis and simulation of numerical schemes for highly-oscillatory nonlinear stochastic Schrödinger equation under diffusion-approximation for instance will be one important objective for us. Finally, an important aspect concerns the quantification of uncertainties in highly-oscillatory kinetic or quantum models (due to an incomplete knowledge of coefficients or imprecise measurement of datas). The construction of efficient multiscale numerical methods which can handle multiple scales as well as random inputs have important engineering applications.

[0] replacing t/ε by $t\omega/\varepsilon$ in (4), with $\omega \in \mathbb{R}^d$ a vector of non-resonant frequencies

MISTIS Project-Team

3. Research Program

3.1. Mixture models

Participants: Alexis Arnaud, Jean-Baptiste Durand, Florence Forbes, Stephane Girard, Julyan Arbel, Daria Bystrova, Giovanni Poggiato, Hongliang Lu, Fabien Boux, Veronica Munoz Ramirez, Benoit Kugler, Alexandre Constantin, Fei Zheng.

Key-words: mixture of distributions, EM algorithm, missing data, conditional independence, statistical pattern recognition, clustering, unsupervised and partially supervised learning.

In a first approach, we consider statistical parametric models, θ being the parameter, possibly multidimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = \{y_1, ..., y_n\}$ and unobserved or missing data $z = \{z_1, ..., z_n\}$. The missing data z_i represents for instance the memberships of one of a set of K alternative categories. The distribution of an observed y_i can be written as a finite mixture of distributions,

$$f(y_i; \theta) = \sum_{k=1}^{K} P(z_i = k; \theta) f(y_i \mid z_i; \theta) \ . \tag{48}$$

These models are interesting in that they may point out hidden variables responsible for most of the observed variability and so that the observed variables are *conditionally* independent. Their estimation is often difficult due to the missing data. The Expectation-Maximization (EM) algorithm is a general and now standard approach to maximization of the likelihood in missing data problems. It provides parameter estimation but also values for missing data.

Mixture models correspond to independent z_i's. They have been increasingly used in statistical pattern recognition. They enable a formal (model-based) approach to (unsupervised) clustering.

3.2. Markov models

Participants: Alexis Arnaud, Brice Olivier, Jean-Baptiste Durand, Florence Forbes, Karina Ashurbekova, Hongliang Lu, Julyan Arbel, Mariia Vladimirova.

Key-words: graphical models, Markov properties, hidden Markov models, clustering, missing data, mixture of distributions, EM algorithm, image analysis, Bayesian inference.

Graphical modelling provides a diagrammatic representation of the dependency structure of a joint probability distribution, in the form of a network or graph depicting the local relations among variables. The graph can have directed or undirected links or edges between the nodes, which represent the individual variables. Associated with the graph are various Markov properties that specify how the graph encodes conditional independence assumptions.

It is the conditional independence assumptions that give graphical models their fundamental modular structure, enabling computation of globally interesting quantities from local specifications. In this way graphical models form an essential basis for our methodologies based on structures.

The graphs can be either directed, e.g. Bayesian Networks, or undirected, e.g. Markov Random Fields. The specificity of Markovian models is that the dependencies between the nodes are limited to the nearest neighbor nodes. The neighborhood definition can vary and be adapted to the problem of interest. When parts of the variables (nodes) are not observed or missing, we refer to these models as Hidden Markov Models (HMM). Hidden Markov chains or hidden Markov fields correspond to cases where the z_i's in (1) are distributed according to a Markov chain or a Markov field. They are a natural extension of mixture models. They are widely used in signal processing (speech recognition, genome sequence analysis) and in image processing (remote sensing, MRI, etc.). Such models are very flexible in practice and can naturally account for the phenomena to be studied.

Hidden Markov models are very useful in modelling spatial dependencies but these dependencies and the possible existence of hidden variables are also responsible for a typically large amount of computation. It follows that the statistical analysis may not be straightforward. Typical issues are related to the neighborhood structure to be chosen when not dictated by the context and the possible high dimensionality of the observations. This also requires a good understanding of the role of each parameter and methods to tune them depending on the goal in mind. Regarding estimation algorithms, they correspond to an energy minimization problem which is NP-hard and usually performed through approximation. We focus on a certain type of methods based on variational approximations and propose effective algorithms which show good performance in practice and for which we also study theoretical properties. We also propose some tools for model selection. Eventually we investigate ways to extend the standard Hidden Markov Field model to increase its modelling power.

3.3. Functional Inference, semi- and non-parametric methods

Participants: Julyan Arbel, Daria Bystrova, Giovanni Poggiato, Stephane Girard, Florence Forbes, Antoine Usseglio Carleve, Pascal Dkengne Sielenou, Meryem Bousebata.

Key-words: dimension reduction, extreme value analysis, functional estimation.

We also consider methods which do not assume a parametric model. The approaches are non-parametric in the sense that they do not require the assumption of a prior model on the unknown quantities. This property is important since, for image applications for instance, it is very difficult to introduce sufficiently general parametric models because of the wide variety of image contents. Projection methods are then a way to decompose the unknown quantity on a set of functions (*e.g.* wavelets). Kernel methods which rely on smoothing the data using a set of kernels (usually probability distributions) are other examples. Relationships exist between these methods and learning techniques using Support Vector Machine (SVM) as this appears in the context of *level-sets estimation* (see section 3.3.2). Such non-parametric methods have become the cornerstone when dealing with functional data [82]. This is the case, for instance, when observations are curves. They enable us to model the data without a discretization step. More generally, these techniques are of great use for *dimension reduction* purposes (section 3.3.3). They enable reduction of the dimension of the functional or multivariate data without assumptions on the observations distribution. Semi-parametric methods refer to methods that include both parametric and non-parametric aspects. Examples include the Sliced Inverse Regression (SIR) method [84] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [81], which is based on the modelling of distribution tails (see section 3.3.1). It differs from traditional statistics which focuses on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distribution tails can be modelled by both a functional part and a real parameter, the extreme value index.

3.3.1. Modelling extremal events

Extreme value theory is a branch of statistics dealing with the extreme deviations from the bulk of probability distributions. More specifically, it focuses on the limiting distributions for the minimum or the maximum of a large collection of random observations from the same arbitrary distribution. Let $X_{1,n} \leq ... \leq X_{n,n}$ denote n ordered observations from a random variable X representing some quantity of interest. A p_n-quantile of X is the value x_{p_n} such that the probability that X is greater than x_{p_n} is p_n, *i.e.*$P(X > x_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation $X_{n,n}$.

To estimate such quantiles therefore requires dedicated methods to extrapolate information beyond the observed values of X. Those methods are based on Extreme value theory. This kind of issue appeared in hydrology. One objective was to assess risk for highly unusual events, such as 100-year floods, starting from flows measured over 50 years. To this end, semi-parametric models of the tail are considered:

$$P(X > x) = x^{-1/\theta}\ell(x),\ x > x_0 > 0, \tag{49}$$

where both the extreme-value index $\theta > 0$ and the function $\ell(x)$ are unknown. The function ℓ is a slowly varying function *i.e.* such that

$$\frac{\ell(tx)}{\ell(x)} \to 1 \text{ as } x \to \infty \tag{50}$$

for all $t > 0$. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developed so far. Such models are often referred to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. It may be necessary to refine the model (2 ,3) by specifying a precise rate of convergence in (3). To this end, a second order condition is introduced involving an additional parameter $\rho \leq 0$. The larger ρ is, the slower the convergence in (3) and the more difficult the estimation of extreme quantiles.
More generally, the problems that we address are part of the risk management theory. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull-tail distributions [10] are defined by their survival distribution function:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{51}$$

Gaussian, gamma, exponential and Weibull distributions, among others, are included in this family. An important part of our work consists in establishing links between models (2) and (4) in order to propose new estimation methods. We also consider the case where the observations were recorded with a covariate information. In this case, the extreme-value index and the p_n-quantile are functions of the covariate. We propose estimators of these functions by using moving window approaches, nearest neighbor methods, or kernel estimators.

3.3.2. Level sets estimation

Level sets estimation is a recurrent problem in statistics which is linked to outlier detection. In biology, one is interested in estimating reference curves, that is to say curves which bound 90% (for example) of the population. Points outside this bound are considered as outliers compared to the reference population. Level sets estimation can be looked at as a conditional quantile estimation problem which benefits from a non-parametric statistical framework. In particular, boundary estimation, arising in image segmentation as well as in supervised learning, is interpreted as an extreme level set estimation problem. Level sets estimation can also be formulated as a linear programming problem. In this context, estimates are sparse since they involve only a small fraction of the dataset, called the set of support vectors.

3.3.3. Dimension reduction

Our work on high dimensional data requires that we face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distortions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [83]. Another field of investigation is to take into account the application goal in the dimension reduction step. One of our approaches is therefore to develop new Gaussian models of high dimensional data for parametric inference [80]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [84].

MODAL Project-Team

3. Research Program

3.1. Research axis 1: Unsupervised learning

Scientific locks related to unsupervised learning are numerous, concerning the clustering outcome validity, the ability to manage different kinds of data, the missing data questioning, the dimensionality of the data set,... Many of them are addressed by the team, leading to publication achievements, often with a specific package delivery (sometimes upgraded as a software or even as a platform grouping several software). Because of the variety of the scope, it involves nearly all the permanent team members, often with PhD students and some engineers. The related works are always embedded inside a probabilistic framework, typically model-based approaches but also model-free ones like PAC-Bayes (PAC stands for Probably Approximately Correct), because such a mathematical environment offers both a well-posed problem and a rigorous answer.

3.2. Research axis 2: Performance assessment

One main concern of the Modal team is to provide theoretical justifications on the procedures which are designed. Such guarantees are important to avoid misleading conclusions resulting from any unsuitable use. For example, one ingredient in proving these guarantees is the use of the PAC framework, leading to finite-sample concentration inequalities. More precisely, contributions to PAC learning rely on the classical empirical process theory and the PAC-Bayesian theory. The Modal team exploits such non-asymptotic tools to analyze the performance of iterative algorithms (such as gradient descent), cross-validation estimators, online change-point detection procedures, ranking algorithms, matrix factorization techniques and clustering methods, for instance. The team also develops some expertise on the formal dynamic study of algorithms related to mixture models (important models used in the previous unsupervised setting), like degeneracy for EM algorithm or also label switching for Gibbs algorithm.

3.3. Research axis 3: Functional data

Mainly due to technological advances, functional data are more and more widespread in many application domains. Functional data analysis (FDA) is concerned with the modeling of data, such as curves, shapes, images or a more complex mathematical object, though as smooth realizations of a stochastic process (an infinite dimensional data object valued in a space of eventually infinite dimension; space of squared integrable functions,...). Time series are an emblematic example even if it should not be limited to them (spectral data, spatial data,...). Basically, FDA considers that data correspond to realizations of stochastic processes, usually assumed to be in a metric, semi-metric, Hilbert or Banach space. One may consider, functional independent or dependent (in time or space) data objects of different types (qualitative, quantitative, ordinal, multivariate, time-dependent, spatial-dependent,...). The last decade saw a dynamic literature on parametric or non-parametric FDA approaches for different types of data and applications to various domains, such as principal component analysis, clustering, regression and prediction.

3.4. Research axis 4: Applications motivating research

The fourth axis consists in translating real application issues into statistical problems raising new (academic) challenges for models developed in Modal team. Cifre Phds in industry and interdisciplinary projects with research teams in Health and Biology are at the core of this objective. The main originality of this objective lies in the use of statistics with complex data, including in particular ultra-high dimension problems. We focus on real applications which cannot be solved by classical data analysis.

MOKAPLAN Project-Team

3. Research Program

3.1. Modeling and Analysis

The first layer of methodological tools developed by our team is a set of theoretical continuous models that aim at formalizing the problems studied in the applications. These theoretical findings will also pave the way to efficient numerical solvers that are detailed in Section 3.2 .

3.1.1. Static Optimal Transport and Generalizations

3.1.1.1. Convexity constraint and Principal Agent problem in Economics.

(*Participants:* G. Carlier, J-D. Benamou, V. Duval, Xavier Dupuis (LUISS Guido Carli University, Roma)) The principal agent problem plays a distinguished role in the literature on asymmetric information and contract theory (with important contributions from several Nobel prizes such as Mirrlees, Myerson or Spence) and it has many important applications in optimal taxation, insurance, nonlinear pricing. The typical problem consists in finding a cost minimizing strategy for a monopolist facing a population of agents who have an unobservable characteristic, the principal therefore has to take into account the so-called incentive compatibilty constraint which is very similar to the cyclical monotonicity condition which characterizes optimal transport plans. In a special case, Rochet and Choné [138] reformulated the problem as a variational problem subject to a convexity constraint. For more general models, and using ideas from Optimal Transportation, Carlier [73] considered the more general c-convexity constraint and proved a general existence result. Using the formulation of [73] McCann, Figalli and Kim [96] gave conditions under which the principal agent problem can be written as an infinite dimensional convex variational problem. The important results of [96] are intimately connected to the regularity theory for optimal transport and showed that there is some hope to numerically solve the principal-agent problem for general utility functions.
Our expertise: We have already contributed to the numerical resolution of the Principal Agent problem in the case of the convexity constraint, see [78], [128], [125].
Goals: So far, the mathematical PA model can be numerically solved for simple utility functions. A Bregman approach inspired by [39] is currently being developed [76] for more general functions. It would be extremely useful as a complement to the theoretical analysis. A new semi-Discrete Geometric approach is also investigated where the method reduces to non-convex polynomial optimization.

3.1.1.2. Optimal transport and conditional constraints in statistics and finance.

(*Participants:* G. Carlier, J-D. Benamou, G. Peyré) A challenging branch of emerging generalizations of Optimal Transportation arising in *economics, statistics and finance* concerns Optimal Transportation with *conditional* constraints. The *martingale optimal transport* [33], [101] which appears naturally in mathematical finance aims at computing robust bounds on option prices as the value of an optimal transport problem where not only the marginals are fixed but the coupling should be the law of a martingale, since it represents the prices of the underlying asset under the risk-neutral probability at the different dates. Note that as soon as more than two dates are involved, we are facing a multimarginal problem.
Our expertise: Our team has a deep expertise on the topic of OT and its generalization, including many already existing collaboration between its members, see for instance [39], [44], [37] for some representative recent collaborative publications.
Goals: This is a non trivial extension of Optimal Transportation theory and MOKAPLAN will develop numerical methods (in the spirit of entropic regularization) to address it. A popular problem in statistics is the so-called quantile regression problem, recently Carlier, Chernozhukov and Galichon [74] used an Optimal Transportation approach to extend quantile regression to several dimensions. In this approach again, not only fixed marginals constraints are present but also constraints on conditional means. As in the martingale Optimal Transportation problem, one has to deal with an extra conditional constraint. The duality approach usually breaks down under such constraints and characterization of optimal couplings is a challenging task both from a theoretical and numerical viewpoint.

3.1.1.3. JKO gradient flows.

(*Participants:* G. Carlier, J-D. Benamou, M. Laborde, Q. Mérigot, V. Duval) The connection between the static and dynamic transportation problems (see Section 2.3) opens the door to many extensions, most notably by leveraging the use of gradient flows in metric spaces. The flow with respect to the transportation distance has been introduced by Jordan-Kindelherer-Otto (JKO) [108] and provides a variational formulation of many linear and non-linear diffusion equations. The prototypical example is the Fokker Planck equation. We will explore this formalism to study new variational problems over probability spaces, and also to derive innovative numerical solvers. The JKO scheme has been very successfully used to study evolution equations that have the structure of a gradient flow in the Wasserstein space. Indeed many important PDEs have this structure: the Fokker-Planck equation (as was first considered by [108]), the porous medium equations, the granular media equation, just to give a few examples. It also finds application in image processing [61]. Figure 4 shows examples of gradient flows.
Our expertise: There is an ongoing collaboration between the team members on the theoretical and numerical analysis of gradient flows.
Goals: We apply and extend our research on JKO numerical methods to treat various extensions:

- Wasserstein gradient flows with a non displacement convex energy (as in the parabolic-elliptic Keller-Segel chemotaxis model [80])
- systems of evolution equations which can be written as gradient flows of some energy on a product space (possibly mixing the Wasserstein and L^2 structures) : multi-species models or the parabolic-parabolic Keller-Segel model [48]
- perturbation of gradient flows: multi-species or kinetic models are not gradient flows, but may be viewed as a perturbation of Wasserstein gradient flows, we shall therefore investigate convergence of splitting methods for such equations or systems.

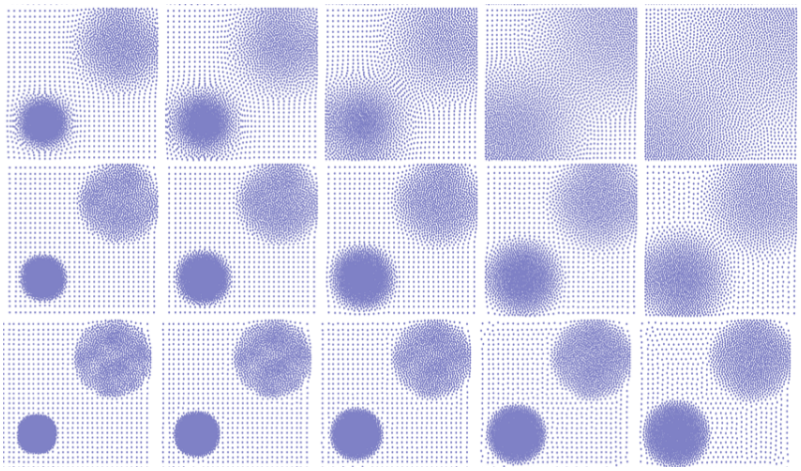


Figure 4. Example of non-linear diffusion equations solved with a JKO flow [40]. The horizontal axis shows the time evolution minimizing the functional $\int \frac{\rho^\alpha}{\alpha-1}$ on the density ρ (discretized here using point clouds, i.e. sum of Diracs' with equal mass). Each row shows a different value of $\alpha = (0.6, 2, 3)$

3.1.1.4. From networks to continuum congestion models.

(*Participants:* G. Carlier, J-D. Benamou, G. Peyré) Congested transport theory in the discrete framework of networks has received a lot of attention since the 50's starting with the seminal work of Wardrop. A few years later, Beckmann proved that equilibria are characterized as solution of a convex minimization problem. However, this minimization problem involves one flow variable per path on the network, its dimension thus quickly becomes too large in practice. An alternative, is to consider continuous in space models of congested optimal transport as was done in [77] which leads to very degenerate PDEs [53].
Our expertise: MOKAPLAN members have contributed a lot to the analysis of congested transport problems and to optimization problems with respect to a metric which can be attacked numerically by fast marching methods [44].
Goals: The case of general networks/anisotropies is still not well understood, general Γ-convergence results will be investigated as well as a detailed analysis of the corresponding PDEs and numerical methods to solve them. Benamou and Carlier already studied numerically some of these PDEs by an augmented Lagrangian method see figure 5 . Note that these class of problems share important similarities with metric learning problem in machine learning, detailed below.

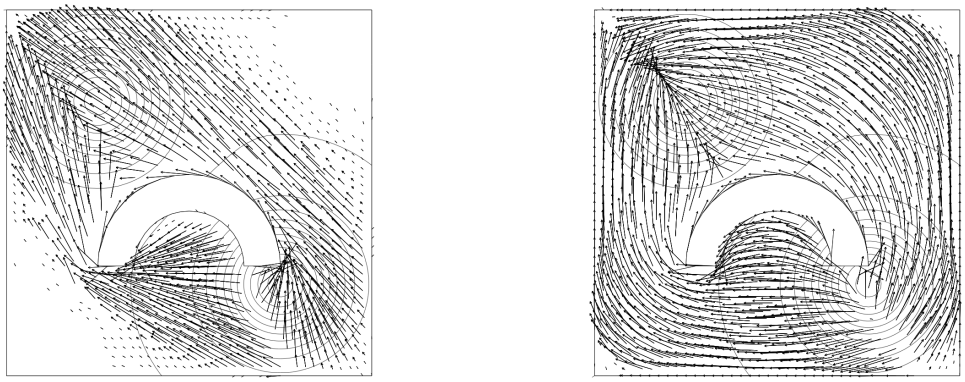


Figure 5. Monge and Wardrop flows of mass around an obstacle [37]. the source/target mass is represented by the level curves. Left : no congestion, Right : congestion.

3.1.2. Diffeomorphisms and Dynamical Transport

3.1.2.1. Growth Models for Dynamical Optimal Transport.

(*Participants:* F-X. Vialard, J-D. Benamou, G. Peyré, L. Chizat) A major issue with the standard dynamical formulation of OT is that it does not allow for variation of mass during the evolution, which is required when tackling medical imaging applications such as tumor growth modeling [64] or tracking elastic organ movements [142]. Previous attempts [119], [135] to introduce a source term in the evolution typically lead to mass teleportation (propagation of mass with infinite speed), which is not always satisfactory.
Our expertise: Our team has already established key contributions both to connect OT to fluid dynamics [35] and to define geodesic metrics on the space of shapes and diffeomorphisms [84].
Goals: Lenaic Chizat's PhD thesis aims at bridging the gap between dynamical OT formulation, and LDDDM diffeomorphisms models (see Section 2.3). This will lead to biologically-plausible evolution models that are both more tractable numerically than LDDM competitors, and benefit from strong theoretical guarantees associated to properties of OT.

3.1.2.2. Mean-field games.

(*Participants:* G. Carlier, J-D. Benamou) The Optimal Transportation Computational Fluid Dynamics (CFD) formulation is a limit case of variational Mean-Field Games (MFGs), a new branch of game theory recently developed by J-M. Lasry and P-L. Lions [112] with an extremely wide range of potential applications [104]. Non-smooth proximal optimization methods used successfully for the Optimal Transportation can be used in the case of deterministic MFGs with singular data and/or potentials [38]. They provide a robust treatment of the positivity constraint on the density of players.
Our expertise: J.-D. Benamou has pioneered with Brenier the CFD approach to Optimal Transportation. Regarding MFGs, on the numerical side, our team has already worked on the use of augmented Lagrangian methods in MFGs [37] and on the analytical side [72] has explored rigorously the optimality system for a singular CFD problem similar to the MFG system.
Goals: We will work on the extension to stochastic MFGs. It leads to non-trivial numerical difficulties already pointed out in [26].

3.1.2.3. Macroscopic Crowd motion, congestion and equilibria.

(*Participants:* G. Carlier, J-D. Benamou, Q. Mérigot, F. Santambrogio (U. Paris-Sud), Y. Achdou (Univ. Paris 7), R. Andreev (Univ. Paris 7)) Many models from PDEs and fluid mechanics have been used to give a description of *people or vehicles moving in a congested environment*. These models have to be classified according to the dimension (1D model are mostly used for cars on traffic networks, while 2-D models are most suitable for pedestrians), to the congestion effects ("soft" congestion standing for the phenomenon where high densities slow down the movement, "hard" congestion for the sudden effects when contacts occur, or a certain threshold is attained), and to the possible rationality of the agents Maury et al [123] recently developed a theory for 2D hard congestion models without rationality, first in a discrete and then in a continuous framework. This model produces a PDE that is difficult to attack with usual PDE methods, but has been successfully studied via Optimal Transportation techniques again related to the JKO gradient flow paradigm. Another possibility to model crowd motion is to use the mean field game approach of Lions and Lasry which limits of Nash equilibria when the number of players is large. This also gives macroscopic models where congestion may appear but this time a global equilibrium strategy is modelled rather than local optimisation by players like in the JKO approach. Numerical methods are starting to be available, see for instance [26], [60].
Our expertise: We have developed numerical methods to tackle both the JKO approach and the MFG approach. The Augmented Lagrangian (proximal) numerical method can actually be applied to both models [37], JKO and deterministic MFGs.
Goals: We want to extend our numerical approach to more realistic congestion model where the speed of agents depends on the density, see Figure 6 for preliminary results. Comparison with different numerical approaches will also be performed inside the ANR ISOTACE. Extension of the Augmented Lagrangian approach to Stochastic MFG will be studied.

3.1.2.4. Diffeomorphic image matching.

(*Participants:* F-X. Vialard, G. Peyré, B. Schmitzer, L. Chizat) Diffeomorphic image registration is widely used in medical image analysis. This class of problems can be seen as the computation of a generalized optimal transport, where the optimal path is a geodesic on a group of diffeomorphisms. The major difference between the two approaches being that optimal transport leads to non smooth optimal maps in general, which is however compulsory in diffeomorphic image matching. In contrast, optimal transport enjoys a convex variational formulation whereas in LDDMM the minimization problem is non convex.
Our expertise: F-X. Vialard is an expert of diffeomorphic image matching (LDDMM) [147], [59], [145]. Our team has already studied flows and geodesics over non-Riemannian shape spaces, which allows for piecewise smooth deformations [84].

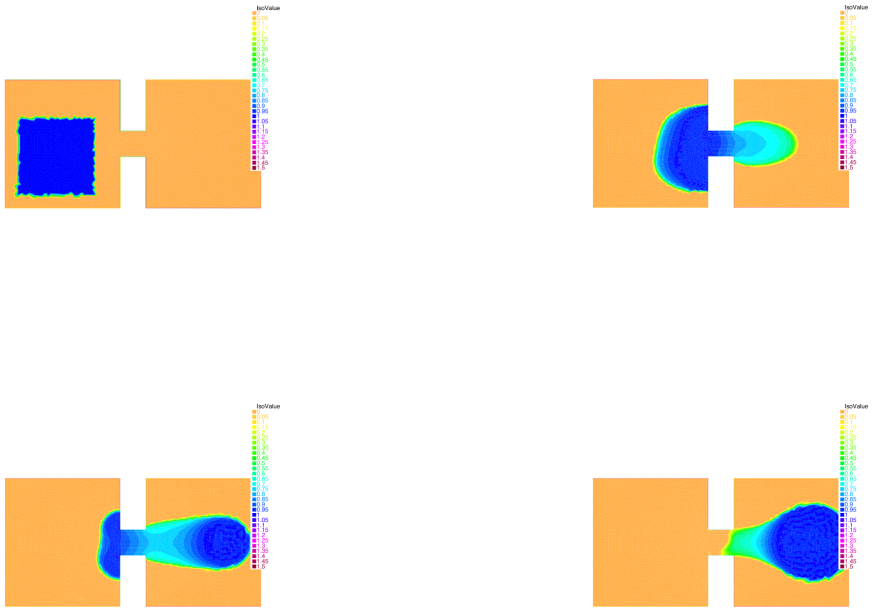


Figure 6. Example of crowd congestion with density dependent speed. The macroscopic density, at 4 different times, of people forced to exit from one room towards a meeting point in a second room.

Goals: Our aim consists in bridging the gap between standard optimal transport and diffeomorphic methods by building new diffeomorphic matching variational formulations that are convex (geometric obstructions might however appear). A related perspective is the development of new registration/transport models in a Lagrangian framework, in the spirit of [141], [142] to obtain more meaningful statistics on longitudinal studies.

Diffeomorphic matching consists in the minimization of a functional that is a sum of a deformation cost and a similarity measure. The choice of the similarity measure is as important as the deformation cost. It is often chosen as a norm on a Hilbert space such as functions, currents or varifolds. From a Bayesian perspective, these similarity measures are related to the noise model on the observed data which is of geometric nature and it is not taken into account when using Hilbert norms. Optimal transport fidelity have been used in the context of signal and image denoising [114], and it is an important question to extends these approach to registration problems. Therefore, we propose to develop similarity measures that are geometric and computationally very efficient using entropic regularization of optimal transport.

Our approach is to use a regularized optimal transport to design new similarity measures on all of those Hilbert spaces. Understanding the precise connections between the evolution of shapes and probability distributions will be investigated to cross-fertilize both fields by developing novel transportation metrics and diffeomorphic shape flows.

The corresponding numerical schemes are however computationally very costly. Leveraging our understanding of the dynamic optimal transport problem and its numerical resolution, we propose to develop new algorithms. These algorithms will use the smoothness of the Riemannian metric to improve both accuracy and speed, using for instance higher order minimization algorithm on (infinite dimensional) manifolds.

3.1.2.5. Metric learning and parallel transport for statistical applications.

(*Participants:* F-X. Vialard, G. Peyré, B. Schmitzer, L. Chizat) The LDDMM framework has been advocated to enable statistics on the space of shapes or images that benefit from the estimation of the deformation. The statistical results of it strongly depend on the choice of the Riemannian metric. A possible direction consists in learning the right invariant Riemannian metric as done in [148] where a correlation matrix (Figure 7) is learnt which represents the covariance matrix of the deformation fields for a given population of shapes. In the same direction, a question of emerging interest in medical imaging is the analysis of time sequence of shapes (called longitudinal analysis) for early diagnosis of disease, for instance [97]. A key question is the inter subject comparison of the organ evolution which is usually done by transport of the time evolution in a common coordinate system via parallel transport or other more basic methods. Once again, the statistical results (Figure 8) strongly depend on the choice of the metric or more generally on the connection that defines parallel transport.
Our expertise: Our team has already studied statistics on longitudinal evolutions in [97], [98].
Goals: Developing higher order numerical schemes for parallel transport (only low order schemes are available at the moment) and developing variational models to learn the metric or the connections for improving statistical results.

3.1.3. Sparsity in Imaging

3.1.3.1. Inverse problems over measures spaces.

(*Participants:* G. Peyré, V. Duval, C. Poon, Q. Denoyelle) As detailed in Section 2.4 , popular methods for regularizing inverse problems in imaging make use of variational analysis over infinite-dimensional (typically non-reflexive) Banach spaces, such as Radon measures or bounded variation functions.
Our expertise: We have recently shown in [146] how – in the finite dimensional case – the non-smoothness of the functionals at stake is crucial to enforce the emergence of geometrical structures (edges in images or fractures in physical materials [49]) for discrete (finite dimensional) problems. We extended this result in a simple infinite dimensional setting, namely sparse regularization of Radon measures for deconvolution [92]. A deep understanding of those continuous inverse problems is crucial to analyze the behavior of their discrete counterparts, and in [93] we have taken advantage of this understanding to develop a fine analysis of the artifacts induced by discrete (*i.e.* which involve grids) deconvolution models. These works are also closely

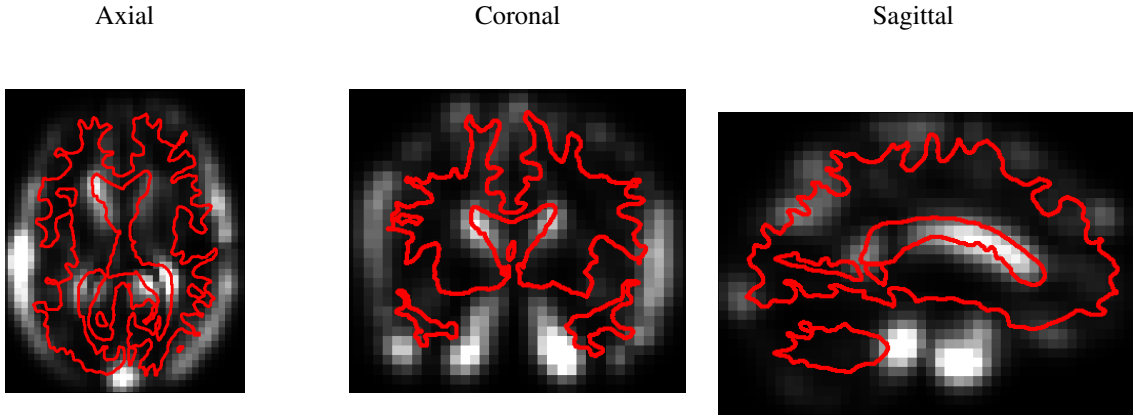


Figure 7. Learning Riemannian metrics in diffeomorphic image matching to capture the brain variability: a diagonal operator that encodes the Riemannian metric is learnt on a template brain out of a collection of brain images. The values of the diagonal operator are shown in greyscale. The red curves represent the boundary between white and grey matter. For more details, we refer the reader to [148], which was a first step towards designing effective and robust metric learning algorithms.

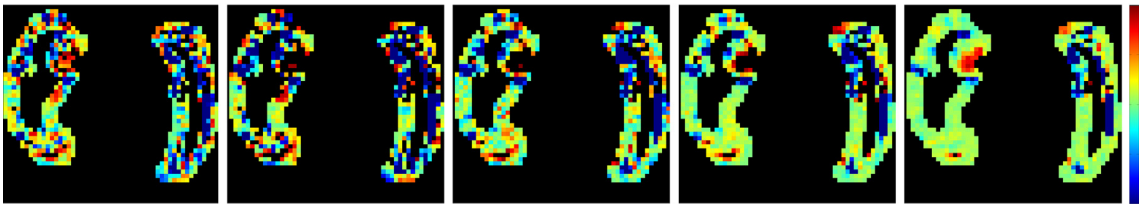


Figure 8. Statistics on initial momenta: In [97], we compared several intersubject transport methodologies to perform statistics on longitudinal evolutions. These longitudinal evolutions are represented by an initial velocity field on the shapes boundaries and these velocity fields are then compared using logistic regression methods that are regularized. The four pictures represent different regularization methods such as L^2, H^1 and regularization including a sparsity prior such as Lasso, Fused Lasso and TV.

related to the problem of limit analysis and yield design in mechanical plasticity, see [75], [49] for an existing collaboration between MOKAPLAN's team members.
Goals: A current major front of research in the mathematical analysis of inverse problems is to extend these results for more complicated infinite dimensional signal and image models, such as for instance the set of piecewise regular functions. The key bottleneck is that, contrary to sparse measures (which are finite sums of Dirac masses), here the objects to recover (smooth edge curves) are not parameterized by a finite number of degrees of freedom. The relevant previous work in this direction are the fundamental results of Chambolle, Caselles and co-workers [34], [28], [81]. They however only deal with the specific case where there is no degradation operator and no noise in the observations. We believe that adapting these approaches using our construction of vanishing derivative pre-certificate [92] could lead to a solution to these theoretical questions.

3.1.3.2. Sub-Riemannian diffusions.

(*Participants:* G. Peyré, J-M. Mirebeau, D. Prandi) Modeling and processing natural images require to take into account their geometry through anisotropic diffusion operators, in order to denoise and enhance directional features such as edges and textures [134], [94]. This requirement is also at the heart of recently proposed models of cortical processing [133]. A mathematical model for these processing is diffusion on sub-Riemanian manifold. These methods assume a fixed, usually linear, mapping from the 2-D image to a lifted function defined on the product of space and orientation (which in turn is equipped with a sub-Riemannian manifold structure).
Our expertise: J-M. Mirebeau is an expert in the discretization of highly anisotropic diffusions through the use of locally adaptive computational stencils [126], [94]. G. Peyré has done several contributions on the definition of geometric wavelets transform and directional texture models, see for instance [134]. Dario Prandi has recently applied methods from sub-Riemannian geometry to image restoration [51].
Goals: A first aspect of this work is to study non-linear, data-adaptive, lifting from the image to the space/orientation domain. This mapping will be implicitly defined as the solution of a convex variational problem. This will open both theoretical questions (existence of a solution and its geometrical properties, when the image to recover is piecewise regular) and numerical ones (how to provide a faithful discretization and fast second order Newton-like solvers). A second aspect of this task is to study the implication of these models for biological vision, in a collaboration with the UNIC Laboratory (directed by Yves Fregnac), located in Gif-sur-Yvette. In particular, the study of the geometry of singular vectors (or "ground states" using the terminology of [45]) of the non-linear sub-Riemannian diffusion operators is highly relevant from a biological modeling point of view.

3.1.3.3. Sparse reconstruction from scanner data.

(*Participants:* G. Peyré, V. Duval, C. Poon) Scanner data acquisition is mathematically modeled as a (sub-sampled) Radon transform [105]. It is a difficult inverse problem because the Radon transform is ill-posed and the set of observations is often aggressively sub-sampled and noisy [140]. Typical approaches [111] try to recover piecewise smooth solutions in order to recover precisely the position of the organ being imaged. There is however a very poor understanding of the actual performance of these methods, and little is known on how to enhance the recovery.
Our expertise: We have obtained a good understanding of the performance of inverse problem regularization on *compact* domains for pointwise sources localization [92].
Goals: We aim at extending the theoretical performance analysis obtained for sparse measures [92] to the set of piecewise regular 2-D and 3-D functions. Some interesting previous work of C. Poon et al [136] (C. Poon is currently a postdoc in MOKAPLAN) have tackled related questions in the field of variable Fourier sampling for compressed sensing application (which is a toy model for fMRI imaging). These approaches are however not directly applicable to Radon sampling, and require some non-trivial adaptations. We also aim at better exploring the connection of these methods with optimal-transport based fidelity terms such as those introduced in [25].

3.1.3.4. Tumor growth modeling in medical image analysis.

(*Participants:* G. Peyré, F-X. Vialard, J-D. Benamou, L. Chizat) Some applications in medical image analysis require to track shapes whose evolution is governed by a growth process. A typical example is tumor growth, where the evolution depends on some typically unknown but meaningful parameters that need to be estimated. There exist well-established mathematical models [64], [132] of non-linear diffusions that take into account recently biologically observed property of tumors. Some related optimal transport models with mass variations have also recently been proposed [121], which are connected to so-called metamorphoses models in the LDDMM framework [46].
Our expertise: Our team has a strong experience on both dynamical optimal transport models and diffeomorphic matching methods (see Section 3.1.2).
Goals: The close connection between tumor growth models [64], [132] and gradient flows for (possibly non-Euclidean) Wasserstein metrics (see Section 3.1.2) makes the application of the numerical methods we develop particularly appealing to tackle large scale forward tumor evolution simulation. A significant departure from the classical OT-based convex models is however required. The final problem we wish to solve is the backward (inverse) problem of estimating tumor parameters from noisy and partial observations. This also requires to set-up a meaningful and robust data fidelity term, which can be for instance a generalized optimal transport metric.

3.2. Numerical Tools

The above continuous models require a careful discretization, so that the fundamental properties of the models are transferred to the discrete setting. Our team aims at developing innovative discretization schemes as well as associated fast numerical solvers, that can deal with the geometric complexity of the variational problems studied in the applications. This will ensure that the discrete solution is correct and converges to the solution of the continuous model within a guaranteed precision. We give below examples for which a careful mathematical analysis of the continuous to discrete model is essential, and where dedicated non-smooth optimization solvers are required.

3.2.1. Geometric Discretization Schemes

3.2.1.1. Discretizing the cone of convex constraints.

(*Participants:* J-D. Benamou, G. Carlier, J-M. Mirebeau, Q. Mérigot) Optimal transportation models as well as continuous models in economics can be formulated as infinite dimensional convex variational problems with the constraint that the solution belongs to the cone of convex functions. Discretizing this constraint is however a tricky problem, and usual finite element discretizations fail to converge.
Our expertise: Our team is currently investigating new discretizations, see in particular the recent proposal [43] for the Monge-Ampère equation and [125] for general non-linear variational problems. Both offer convergence guarantees and are amenable to fast numerical resolution techniques such as Newton solvers. Since [43] explaining how to treat efficiently and in full generality Transport Boundary Conditions for Monge-Ampère, this is a promising fast and new approach to compute Optimal Transportation viscosity solutions. A monotone scheme is needed. One is based on Froese Oberman work [100], a new different and more accurate approach has been proposed by Mirebeau, Benamou and Collino [41]. As shown in [86], discretizing the constraint for a continuous function to be convex is not trivial. Our group has largely contributed to solve this problem with G. Carlier [78], Quentin Mérigot [128] and J-M. Mirebeau [125]. This problem is connected to the construction of monotone schemes for the Monge-Ampère equation.
Goals: The current available methods are 2-D. They need to be optimized and parallelized. A non-trivial extension to 3-D is necessary for many applications. The notion of c-convexity appears in optimal transport for generalized displacement costs. How to construct an adapted discretization with "good" numerical properties is however an open problem.

3.2.1.2. Numerical JKO gradient flows.

(*Participants:* J-D. Benamou, G. Carlier, J-M. Mirebeau, G. Peyré, Q. Mérigot) As detailed in Section 2.3 , gradient Flows for the Wasserstein metric (aka JKO gradient flows [108]) provides a variational formulation of many non-linear diffusion equations. They also open the way to novel discretization schemes. From a computational point, although the JKO scheme is constructive (it is based on the implicit Euler scheme), it has not been very much used in practice numerically because the Wasserstein term is difficult to handle (except in dimension one).
Our expertise:

Solving one step of a JKO gradient flow is similar to solving an Optimal transport problem. A geometrical a discretization of the Monge-Ampère operator approach has been proposed by Mérigot, Carlier, Oudet and Benamou in [40] see Figure 4 . The Gamma convergence of the discretisation (in space) has been proved.
Goals: We are also investigating the application of other numerical approaches to Optimal Transport to JKO gradient flows either based on the CFD formulation or on the entropic regularization of the Monge-Kantorovich problem (see section 3.2.3). An in-depth study and comparison of all these methods will be necessary.

3.2.2. Sparse Discretization and Optimization

3.2.2.1. From discrete to continuous sparse regularization and transport.

(*Participants:* V. Duval, G. Peyré, G. Carlier, Jalal Fadili (ENSICaen), Jérôme Malick (CNRS, Univ. Grenoble)) While pervasive in the numerical analysis community, the problem of discretization and Γ-convergence from discrete to continuous is surprisingly over-looked in imaging sciences. To the best of our knowledge, our recent work [92], [93] is the first to give a rigorous answer to the transition from discrete to continuous in the case of the spike deconvolution problem. Similar problems of Γ-convergence are progressively being investigated in the optimal transport community, see in particular [79].
Our expertise: We have provided the first results on the discrete-to-continous convergence in both sparse regularization variational problems [92], [93] and the static formulation of OT and Wasserstein barycenters [79]
Goals: In a collaboration with Jérôme Malick (Inria Grenoble), our first goal is to generalize the result of [92] to generic partly-smooth convex regularizers routinely used in imaging science and machine learning, a prototypal example being the nuclear norm (see [146] for a review of this class of functionals). Our second goal is to extend the results of [79] to the novel class of entropic discretization schemes we have proposed [39], to lay out the theoretical foundation of these ground-breaking numerical schemes.

3.2.2.2. Polynomial optimization for grid-free regularization.

(*Participants:* G. Peyré, V. Duval, I. Waldspurger) There has been a recent spark of attention of the imaging community on so-called "grid free" methods, where one tries to directly tackle the infinite dimensional recovery problem over the space of measures, see for instance [71], [92]. The general idea is that if the range of the imaging operator is finite dimensional, the associated dual optimization problem is also finite dimensional (for deconvolution, it corresponds to optimization over the set of trigonometric polynomials).
Our expertise: We have provided in [92] a sharp analysis of the support recovery property of this class of methods for the case of sparse spikes deconvolution.
Goals: A key bottleneck of these approaches is that, while being finite dimensional, the dual problem necessitates to handle a constraint of polynomial positivity, which is notoriously difficult to manipulate (except in the very particular case of 1-D problems, which is the one exposed in [71]). A possible, but very costly, methodology is to ressort to Lasserre's SDP representation hierarchy [113]. We will make use of these approaches and study how restricting the level of the hierarchy (to obtain fast algorithms) impacts the recovery performances (since this corresponds to only computing approximate solutions). We will pay a particular attention to the recovery of 2-D piecewise constant functions (the so-called total variation of functions regularization [139]), see Figure 3 for some illustrative applications of this method.

3.2.3. First Order Proximal Schemes

3.2.3.1. L^2 proximal methods.

(*Participants:* G. Peyré, J-D. Benamou, G. Carlier, Jalal Fadili (ENSICaen)) Both sparse regularization problems in imaging (see Section 2.4) and dynamical optimal transport (see Section 2.3) are instances of large scale, highly structured, non-smooth convex optimization problems. First order proximal splitting optimization algorithms have recently gained lots of interest for these applications because they are the only ones capable of scaling to giga-pixel discretizations of images and volumes and at the same time handling non-smooth objective functions. They have been successfully applied to optimal transport [35], [129], congested optimal transport [63] and to sparse regularizations (see for instance [137] and the references therein).
Our expertise: The pioneering work of our team has shown how these proximal solvers can be used to tackle the dynamical optimal transport problem [35], see also [129]. We have also recently developed new proximal schemes that can cope with non-smooth composite objectives functions [137].
Goals: We aim at extending these solvers to a wider class of variational problems, most notably optimization under divergence constraints [37]. Another subject we are investigating is the extension of these solvers to both non-smooth and non-convex objective functionals, which are mandatory to handle more general transportation problems and novel imaging regularization penalties.

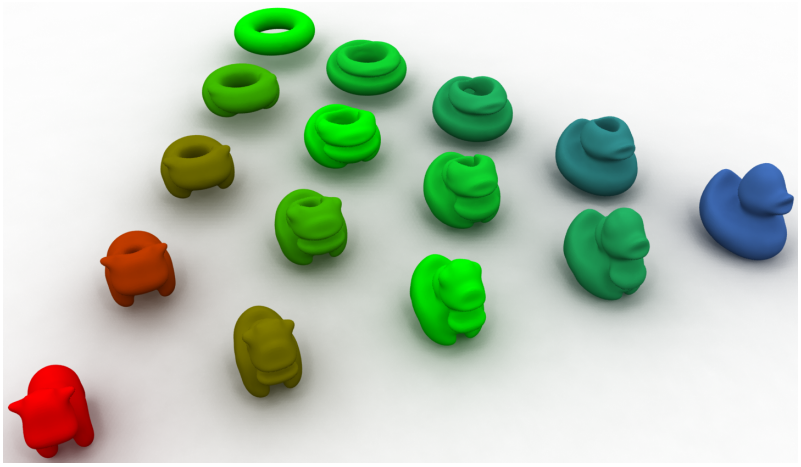


Figure 9. Example of barycenter between shapes computed using optimal transport barycenters of the uniform densities inside the 3 extremal shapes, computed as detailed in [143]. Note that the barycenters are not in general uniform distributions, and we display them as the surface defined by a suitable level-set of the density.

3.2.3.2. Bregman proximal methods.

(*Participants:* G. Peyré G. Carlier, L. Nenna, J-D. Benamou, L. Nenna, Marco Cuturi (Kyoto Univ.)) The entropic regularization of the Kantorovich linear program for OT has been shown to be surprisingly simple and efficient, in particular for applications in machine learning [90]. As shown in [39], this is a special instance of the general method of Bregman iterations, which is also a particular instance of first order proximal schemes according to the Kullback-Leibler divergence.

Our expertise: We have recently [39] shown how Bregman projections [54] and Dykstra algorithm [31] offer a generic optimization framework to solve a variety of generalized OT problems. Carlier and Dupuis [76] have designed a new method based on alternate Dykstra projections and applied it to the *principal-agent problem* in microeconomics. We have applied this method in computer graphics in a paper accepted in SIGGRAPH 2015 [143]. Figure 9 shows the potential of our approach to handle giga-voxel datasets: the input volumetric densities are discretized on a 100^3 computational grid.
Goals: Following some recent works (see in particular [83]) we first aim at studying primal-dual optimization schemes according to Bregman divergences (that would go much beyond gradient descent and iterative projections), in order to offer a versatile and very effective framework to solve variational problems involving OT terms. We then also aim at extending the scope of usage of this method to applications in quantum mechanics (Density Functional Theory, see [87]) and fluid dynamics (Brenier's weak solutions of the incompressible Euler equation, see [55]). The computational challenge is that realistic physical examples are of a huge size not only because of the space discretization of one marginal but also because of the large number of marginals involved (for incompressible Euler the number of marginals equals the number of time steps).

NACHOS Project-Team

3. Research Program

3.1. Scientific foundations

The research activities undertaken by the team aim at developing innovative numerical methodologies putting the emphasis on several features:

- **Accuracy**. The foreseen numerical methods should rely on discretization techniques that best fit to the geometrical characteristics of the problems at hand. Methods based on unstructured, locally refined, even non-conforming, simplicial meshes are particularly attractive in this regard. In addition, the proposed numerical methods should also be capable to accurately describe the underlying physical phenomena that may involve highly variable space and time scales. Both objectives are generally addressed by studying so-called hp-adaptive solution strategies which combine h-adaptivity using local refinement/coarsening of the mesh and p-adaptivity using adaptive local variation of the interpolation order for approximating the solution variables. However, for physical problems involving strongly heterogeneous or high contrast propagation media, such a solution strategy may not be sufficient. Then, for dealing accurately with these situations, one has to design numerical methods that specifically address the multiscale nature of the underlying physical phenomena.
- **Numerical efficiency**. The simulation of unsteady problems most often relies on explicit time integration schemes. Such schemes are constrained by a stability criterion, linking some space and time discretization parameters, that can be very restrictive when the underlying mesh is highly non-uniform (especially for locally refined meshes). For realistic 3D problems, this can represent a severe limitation with regards to the overall computing time. One possible overcoming solution consists in resorting to an implicit time scheme in regions of the computational domain where the underlying mesh size is very small, while an explicit time scheme is applied elsewhere in the computational domain. The resulting hybrid explicit-implicit time integration strategy raises several challenging questions concerning both the mathematical analysis (stability and accuracy, especially for what concern numerical dispersion), and the computer implementation on modern high performance systems (data structures, parallel computing aspects). A second, often considered approach is to devise a local time stepping strategy. Beside, when considering time-harmonic (frequency-domain) wave propagation problems, numerical efficiency is mainly linked to the solution of the system of algebraic equations resulting from the discretization in space of the underlying PDE model. Various strategies exist ranging from the more robust and efficient sparse direct solvers to the more flexible and cheaper (in terms of memory resources) iterative methods. Current trends tend to show that the ideal candidate will be a judicious mix of both approaches by relying on domain decomposition principles.
- **Computational efficiency**. Realistic 3D wave propagation problems involve the processing of very large volumes of data. The latter results from two combined parameters: the size of the mesh i.e the number of mesh elements, and the number of degrees of freedom per mesh element which is itself linked to the degree of interpolation and to the number of physical variables (for systems of partial differential equations). Hence, numerical methods must be adapted to the characteristics of modern parallel computing platforms taking into account their hierarchical nature (e.g multiple processors and multiple core systems with complex cache and memory hierarchies). In addition, appropriate parallelization strategies need to be designed that combine SIMD and MIMD programming paradigms.

From the methodological point of view, the research activities of the team are concerned with four main topics: (1) high order finite element type methods on unstructured or hybrid structured/unstructured meshes for the discretization of the considered systems of PDEs, (2) efficient time integration strategies for dealing with grid induced stiffness when using non-uniform (locally refined) meshes, (3) numerical treatment of complex propagation media models (e.g. physical dispersion models), (4) algorithmic adaptation to modern high performance computing platforms.

3.2. High order discretization methods

3.2.1. The Discontinuous Galerkin method

The Discontinuous Galerkin method (DG) was introduced in 1973 by Reed and Hill to solve the neutron transport equation. From this time to the 90's a review on the DG methods would likely fit into one page. In the meantime, the Finite Volume approach (FV) has been widely adopted by computational fluid dynamics scientists and has now nearly supplanted classical finite difference and finite element methods in solving problems of non-linear convection and conservation law systems. The success of the FV method is due to its ability to capture discontinuous solutions which may occur when solving non-linear equations or more simply, when convecting discontinuous initial data in the linear case. Let us first remark that DG methods share with FV methods this property since a first order FV scheme may be viewed as a 0th order DG scheme. However a DG method may also be considered as a Finite Element (FE) one where the continuity constraint at an element interface is released. While keeping almost all the advantages of the FE method (large spectrum of applications, complex geometries, etc.), the DG method has other nice properties which explain the renewed interest it gains in various domains in scientific computing as witnessed by books or special issues of journals dedicated to this method [47]- [48]- [49]- [54]:

- It is naturally adapted to a high order approximation of the unknown field. Moreover, one may increase the degree of the approximation in the whole mesh as easily as for spectral methods but, with a DG method, this can also be done very locally. In most cases, the approximation relies on a polynomial interpolation method but the DG method also offers the flexibility of applying local approximation strategies that best fit to the intrinsic features of the modeled physical phenomena.
- When the space discretization is coupled to an explicit time integration scheme, the DG method leads to a block diagonal mass matrix whatever the form of the local approximation (e.g. the type of polynomial interpolation). This is a striking difference with classical, continuous FE formulations. Moreover, the mass matrix may be diagonal if the basis functions are orthogonal.
- It easily handles complex meshes. The grid may be a classical conforming FE mesh, a non-conforming one or even a hybrid mesh made of various elements (tetrahedra, prisms, hexahedra, etc.). The DG method has been proven to work well with highly locally refined meshes. This property makes the DG method more suitable (and flexible) to the design of some hp-adaptive solution strategy.
- It is also flexible with regards to the choice of the time stepping scheme. One may combine the DG spatial discretization with any global or local explicit time integration scheme, or even implicit, provided the resulting scheme is stable.
- It is naturally adapted to parallel computing. As long as an explicit time integration scheme is used, the DG method is easily parallelized. Moreover, the compact nature of DG discretization schemes is in favor of high computation to communication ratio especially when the interpolation order is increased.

As with standard FE methods, a DG method relies on a variational formulation of the continuous problem at hand. However, due to the discontinuity of the global approximation, this variational formulation has to be defined locally, at the element level. Then, a degree of freedom in the design of a DG method stems from the approximation of the boundary integral term resulting from the application of an integration by parts to the element-wise variational form. In the spirit of FV methods, the approximation of this boundary integral term calls for a numerical flux function which can be based on either a centered scheme or an upwind scheme, or a blending between these two schemes.

3.2.2. High order DG methods for wave propagation models

DG methods are at the heart of the activities of the team regarding the development of high order discretization schemes for the PDE systems modeling electromagnetic and elatsodynamic wave propagation.

- **Nodal DG methods for time-domain problems**. For the numerical solution of the time-domain Maxwell equations, we have first proposed a non-dissipative high order DGTD (Discontinuous Galerkin Time-Domain) method working on unstructured conforming simplicial meshes [9]. This DG method combines a central numerical flux function for the approximation of the integral term at the interface of two neighboring elements with a second order leap-frog time integration scheme. Moreover, the local approximation of the electromagnetic field relies on a nodal (Lagrange type) polynomial interpolation method. Recent achievements by the team deal with the extension of these methods towards non-conforming unstructured [6]-[7] and hybrid structured/unstructured meshes [4], their coupling with hybrid explicit/implicit time integration schemes in order to improve their efficiency in the context of locally refined meshes [3]-[14]-[13]. A high order DG method has also been proposed for the numerical resolution of the elastodynamic equations modeling the propagation of seismic waves [2].
- **Hybridizable DG (HDG) method for time-domain and time-harmonic problems**. For the numerical treatment of the time-harmonic Maxwell equations, nodal DG methods can also be considered [5]. However, such DG formulations are highly expensive, especially for the discretization of 3D problems, because they lead to a large sparse and undefinite linear system of equations coupling all the degrees of freedom of the unknown physical fields. Different attempts have been made in the recent past to improve this situation and one promising strategy has been recently proposed by Cockburn *et al.*[52] in the form of so-called hybridizable DG formulations. The distinctive feature of these methods is that the only globally coupled degrees of freedom are those of an approximation of the solution defined only on the boundaries of the elements. This work is concerned with the study of such Hybridizable Discontinuous Galerkin (HDG) methods for the solution of the system of Maxwell equations in the time-domain when the time integration relies on an implicit scheme, or in the frequency-domain. The team has been a precursor in the development of HDG methods for the frequency-domain Maxwell equations[12].
- **Multiscale DG methods for time-domain problems**. More recently, in collaboration with LNCC in Petropolis (Frédéric Valentin) the framework of the HOMAR assoacite team, we are investigating a family of methods specifically designed for an accurate and efficient numerical treatment of multiscale wave propagation problems. These methods, referred to as Multiscale Hybrid Mixed (MHM) methods, are currently studied in the team for both time-domain electromagnetic and elastodynamic PDE models. They consist in reformulating the mixed variational form of each system into a global (arbitrarily coarse) problem related to a weak formulation of the boundary condition (carried by a Lagrange multiplier that represents e.g. the normal stress tensor in elastodynamic sytems), and a series of small, element-wise, fully decoupled problems resembling to the initial one and related to some well chosen partition of the solution variables on each element. By construction, that methodology is fully parallelizable and recursivity may be used in each local problem as well, making MHM methods belonging to multi-level highly parallelizable methods. Each local problem may be solved using DG or classical Galerkin FE approximations combined with some appropriate time integration scheme (θ-scheme or leap-frog scheme).

3.3. Efficient time integration strategies

The use of unstructured meshes (based on triangles in two space dimensions and tetrahedra in three space dimensions) is an important feature of the DGTD methods developed in the team which can thus easily deal with complex geometries and heterogeneous propagation media. Moreover, DG discretization methods are naturally adapted to local, conforming as well as non-conforming, refinement of the underlying mesh. Most of the existing DGTD methods rely on explicit time integration schemes and lead to block diagonal mass matrices which is often recognized as one of the main advantages with regards to continuous finite element methods.

However, explicit DGTD methods are also constrained by a stability condition that can be very restrictive on highly refined meshes and when the local approximation relies on high order polynomial interpolation. There are basically three strategies that can be considered to cure this computational efficiency problem. The first approach is to use an unconditionally stable implicit time integration scheme to overcome the restrictive constraint on the time step for locally refined meshes. In a second approach, a local time stepping strategy is combined with an explicit time integration scheme. In the third approach, the time step size restriction is overcome by using a hybrid explicit-implicit procedure. In this case, one blends a time implicit and a time explicit schemes where only the solution variables defined on the smallest elements are treated implicitly. The first and third options are considered in the team in the framework of DG [3]-[14]-[13] and HDG discretization methods.

3.4. Numerical treatment of complex material models

Towards the general aim of being able to consider concrete physical situations, we are interested in taking into account in the numerical methodologies that we study, a better description of the propagation of waves in realistic media. In the case of electromagnetics, a typical physical phenomenon that one has to consider is *dispersion*. It is present in almost all media and expresses the way the material reacts to an electromagnetic field. In the presence of an electric field a medium does not react instantaneously and thus presents an electric polarization of the molecules or electrons that itself influences the electric displacement. In the case of a linear homogeneous isotropic media, there is a linear relation between the applied electric field and the polarization. However, above some range of frequencies (depending on the considered material), the dispersion phenomenon cannot be neglected and the relation between the polarization and the applied electric field becomes complex. This is rendered via a frequency-dependent complex permittivity. Several models of complex permittivity exist. Concerning biological media, the Debye model is commonly adopted in the presence of water, biological tissues and polymers, so that it already covers a wide range of applications [11]. In the context of nanoplasmonics, one is interested in modeling the dispersion effects on metals on the nanometer scale and at optical frequencies. In this case, the Drude or the Drude-Lorentz models are generally chosen [17]. In the context of seismic wave propagation, we are interested by the intrinsic attenuation of the medium [15]. In realistic configurations, for instance in sedimentary basins where the waves are trapped, we can observe site effects due to local geological and geotechnical conditions which result in a strong increase in amplification and duration of the ground motion at some particular locations. During the wave propagation in such media, a part of the seismic energy is dissipated because of anelastic losses relied to the internal friction of the medium. For these reasons, numerical simulations based on the basic assumption of linear elasticity are no more valid since this assumption results in a severe overestimation of amplitude and duration of the ground motion, even when we are not in presence of a site effect, since intrinsic attenuation is not taken into account.

3.5. High performance numerical computing

Beside basic research activities related to the design of numerical methods and resolution algorithms for the wave propagation models at hand, the team is also committed to demonstrate the benefits of the proposed numerical methodologies in the simulation of challenging three-dimensional problems pertaining to computational electromagnetics and computational geoseismics. For such applications, parallel computing is a mandatory path. Nowadays, modern parallel computers most often take the form of clusters of heterogeneous multiprocessor systems, combining multiple core CPUs with accelerator cards (e.g Graphical Processing Units - GPUs), with complex hierarchical distributed-shared memory systems. Developing numerical algorithms that efficiently exploit such high performance computing architectures raises several challenges, especially in the context of a massive parallelism. In this context, current efforts of the team are towards the exploitation of multiple levels of parallelism (computing systems combining CPUs and GPUs) through the study of hierarchical SPMD (Single Program Multiple Data) strategies for the parallelization of unstructured mesh based solvers.

NANO-D Team

3. Research Program

3.1. The need for practical design of nanosystems

Computing has long been an essential tool of engineering. During the twentieth century, the development of macroscopic engineering has been largely stimulated by progress in numerical design and prototyping. Cars, planes, boats, and many other manufactured objects are nowadays, for the most part, designed and tested on computers. Digital prototypes have progressively replaced actual ones, and effective computer-aided engineering tools (e.g., CATIA, SolidWorks, T-FLEX CAD, Alibre Design, TopSolid, etc.) have helped cut costs and reduce production cycles of macroscopic systems [61].

The twenty-first century is most likely to see a similar development at the atomic scale. Indeed, the recent years have seen tremendous progress in nanotechnology. The magazine Science, for example, recently featured a paper demonstrating an example of DNA nanotechnology, where DNA strands are stacked together through programmable self-assembly [32]. In February 2007, the cover of Nature Nanotechnology showed a "nano-wheel" composed of a few atoms only. Several nanosystems have already been demonstrated, including a *de-novo* computationally designed protein interface [33], a wheelbarrow molecule [43], a nano-car [65], a Morse molecule [16], etc. Typically, these designs are optimized using semi-empirical quantum mechanics calculations, such as the semi-empirical ASED+ calculation technique [17].

While impressive, these are but two examples of the nanoscience revolution already impacting numerous fields, including electronics and semiconductors [50], textiles [48], [38], energy [52], food [27], drug delivery [37], [68], chemicals [39], materials [28], the automotive industry [14], aerospace and defense [34], medical devices and therapeutics [30], medical diagnostics [69], etc. According to some estimates, the world market for nanotechnology-related products and services will reach one trillion dollars by 2015 [60]. Nano-engineering groups are multiplying throughout the world, both in academia and in the industry: in the USA, the MIT has a "NanoEngineering" research group, Sandia National Laboratories created a "National Institute for Nano Engineering", to name a few; China founded a "National Center for Nano Engineering" in 2003, etc. Europe is also a significant force in public funding of nanoscience and nanotechnology and, in Europe, Grenoble and the Rhone-Alpes area gather numerous institutions and organizations related to nanoscience.

Of course, not all small systems that currently fall under the label "nano" have mechanical, electronic, optical properties similar to the examples given above. Furthermore, current construction capabilities lack behind some of the theoretical designs which have been proposed, such as the planetary gear designed by Eric Drexler at Nanorex. However, the trend is clearly for adding more and more functionality to nanosystems. While designing nanosystems is still very much an art mostly performed by physicists, chemists and biologists in labs throughout the world, there is absolutely no doubt that fundamental engineering practices will progressively emerge, and that these practices will be turned into quantitative rules and methods. Similar to what has happened with macroscopic engineering, powerful and generic software will then be employed to engineer complex nanosystems.

3.2. Challenges of practical nanosystem design

As with macrosystems, designing nanosystems will involve modeling and simulation within software applications: modeling, especially structural modeling, will be concerned with the creation of potentially complex chemical structures such as the examples above, using a graphical user interface, parsers, scripts, builders, etc.; simulation will be employed to predict some properties of the constructed models, including mechanical properties, electronic properties, chemical properties, etc.

In general, design may be considered as an "inverse simulation problem". Indeed, designed systems often need to be optimized so that their properties — predicted by simulation — satisfy specific objectives and constraints (e.g. a car should have a low drag coefficient, a drug should have a high affinity and selectivity to a target protein, a nano-wheel should roll when pushed, etc.). Being the main technique employed to predict properties, simulation is essential to the design process. At the nanoscale, simulation is even more important. Indeed, physics significantly constrains atomic structures (e.g. arbitrary inter-atomic distances cannot exist), so that a tentative atomic shape should be checked for plausibility much earlier in the design process (e.g. remove atomic clashes, prevent unrealistic, high-energy configurations, etc.). For nanosystems, thus, efficient simulation algorithms are required both when modeling structures and when predicting systems properties. Precisely, an effective software tool to design nanosystems should (a) allow for interactive physically-based modeling, where all user actions (e.g. displacing atoms, modifying the system's topology, etc.) are automatically followed by a few steps of energy minimization to help the user build plausible structures, even for large number of atoms, and (b) be able to predict systems properties, through a series of increasingly complex simulations.

3.3. Current simulation approaches

Even though the growing need for effective nanosystem design will still increase the demand for simulation, a lot of research has already gone into the development of efficient simulation algorithms. Typically, two approaches are used: (a) increasing the computational resources (use super-computers, computer clusters, grids, develop parallel computing approaches, etc.), or (b) simulating simplified physics and/or models. Even though the first strategy is sometimes favored, it is expensive and, it could be argued, inefficient: only a few supercomputers exist, not everyone is willing to share idle time from their personal computer, etc. Surely, we would see much less creativity in cars, planes, and manufactured objects all around if they had to be designed on one of these scarce super-resources.

The second strategy has received a lot of attention. Typical approaches to speed up molecular mechanics simulation include lattice simulations [71], removing some degrees of freedom (e.g. keeping torsion angles only [46], [66]), coarse-graining [70], [63], [18], [64], multiple time step methods [57], [58], fast multipole methods [31], parallelization [45], averaging [26], multi-scale modeling [25], [22], reactive force fields [24], [74], interactive multiplayer games for predicting protein structures [29], etc. Until recently, quantum mechanics methods, as well as mixed quantum / molecular mechanics methods were still extremely slow. One breakthrough has consisted in the discovery of linear-scaling, divide-and-conquer quantum mechanics methods [72], [73].

Overall, the computational community has already produced a variety of sophisticated simulation packages, for both classical and quantum simulation: ABINIT, AMBER, CHARMM, Desmond, GROMOS and GROMACS, LAMMPS, NAMD, ROSETTA, SIESTA, TINKER, VASP, YASARA, etc. Some of these tools are open source, while some others are available commercially, sometimes via integrating applications: Ascalaph Designer, BOSS, Discovery Studio, Materials Studio, Maestro, MedeA, MOE, NanoEngineer-1, Spartan, etc. Other tools are mostly concerned with visualization, but may sometimes be connected to simulation packages: Avogadro, PyMol, VMD, Zodiac, etc. The nanoHUB network also includes a rich set of tools related to computational nanoscience.

To the best of our knowledge, however, all methods which attempt to speed up dynamics simulations perform a priori simplification assumptions, which might bias the study of the simulated phenomenon. A few recent, interesting approaches have managed to combine several levels of description (e.g. atomistic and coarse-grained) into a single simulation, and have molecules switch between levels during simulation, including the adaptive resolution method [53], [54], [55], [56], the adaptive multiscale method [51], and the adaptive partitioning of the Lagrangian method [40]. Although these approaches have demonstrated some convincing applications, they all suffer from a number of limitations stemming from the fact that they are either ad hoc methods tuned to fix specific problems (e.g. fix density problems in regions where the level of description changes), or mathematically founded methods that necessitate to "calibrate" potentials so that they can be mixed (i.e. all potentials have to agree on a reference point). In general, multi-scale methods, even when

they do not allow molecules to switch between levels of detail during simulation, have to solve the problem of rigorously combining multiple levels of description (i.e. preserve statistics, etc.), of assigning appropriate levels to different parts of the simulated system ("simplify as much as possible, but not too much"), and of determining computable mappings between levels of description (especially, adding back detail when going from coarse-grained descriptions to fine-grained descriptions).

NECS Team

3. Research Program

3.1. Introduction

NECS team deals with Networked Control Systems. Since its foundation in 2007, the team has been addressing issues of control under imperfections and constraints deriving from the network (limited computation resources of the embedded systems, delays and errors due to communication, limited energy resources), proposing co-design strategies. The team has recently moved its focus towards general problems on *control of network systems*, which involve the analysis and control of dynamical systems with a network structure or whose operation is supported by networks. This is a research domain with substantial growth and is now recognized as a priority sector by the IEEE Control Systems Society: IEEE has started a new journal, IEEE Transactions on Control of Network Systems, whose first issue appeared in 2014.

More in detail, the research program of NECS team is along lines described in the following sections.

3.2. Distributed estimation and data fusion in network systems

This research topic concerns distributed data combination from multiple sources (sensors) and related information fusion, to achieve more specific inference than could be achieved by using a single source (sensor). It plays an essential role in many networked applications, such as communication, networked control, monitoring, navigation and surveillance. Distributed estimation has already been considered in the team. We wish to capitalize and strengthen these activities by focusing on integration of heterogeneous, multidimensional, and large data sets:

- Heterogeneity and large data sets. This issue constitutes a clearly identified challenge for the future. Indeed, heterogeneity comes from the fact that data are given in many forms, refer to different scales, and carry different information. Therefore, data fusion and integration will be achieved by developing new multi-perception mathematical models that can allow tracking continuous (macroscopic) and discrete (microscopic) dynamics under a unified framework while making different scales interact with each other. More precisely, many scales are considered at the same time, and they evolve following a unique fully-integrated dynamics generated by the interactions of the scales. The new multi-perception models will be integrated to forecast, estimate and broadcast useful system states in a distributed way. Targeted applications include traffic networks and navigation.
- Multidimensionality. This issue concerns the analysis and the processing of multidimensional data, organized in multiway array, in a distributed way. Robustness of previously-developed algorithms will be studied. In particular, the issue of missing data will be taken into account. In addition, since the considered multidimensional data are generated by dynamic systems, dynamic analysis of multiway array (or tensors) will be considered. The targeted applications concern distributed detection in complex networks and distributed signal processing for collaborative networks. This topic is developed in strong collaboration with UFC (Brazil).

3.3. Network systems and graph analysis

This is a research topic at the boundaries between graph theory and dynamical systems theory.

A first main line of research will be to study complex systems whose interactions are modeled with graphs, and to unveil the effect of the graph topology on system-theoretic properties such as observability or controllability. In particular, on-going work concerns observability of graph-based systems: after preliminary results concerning consensus systems over distance-regular graphs, the aim is to extend results to more general networks. A special focus will be on the notion of 'generic properties', namely properties which depend only on the underlying graph describing the sparsity pattern, and hold true almost surely with a random choice of the non-zero coefficients. Further work will be to explore situations in which there is the need for new notions different from the classical observability or controllability. For example, in opinion-forming in social networks or in formation of birds flocks, the potential leader might have a goal different from classical controllability. On the one hand, his goal might be much less ambitious than the classical one of driving the system to any possible state (e.g., he might want to drive everybody near its own opinion, only, and not to any combination of different individual opinions), and on the other hand he might have much weaker tools to construct his control input (e.g., he might not know the whole system's dynamics, but only some local partial information). Another example is the question of detectability of an unknown input under the assumption that such an input has a sparsity constraint, a question arising from the fact that a cyber-physical attack might be modeled as an input aiming at controlling the system's state, and that limitations in the capabilities of the attacker might be modeled as a sparsity constraint on the input.

A second line of research will concern graph discovery, namely algorithms aiming at reconstructing some properties of the graph (such as the number of vertices, the diameter, the degree distribution, or spectral properties such as the eigenvalues of the graph Laplacian), using some measurements of quantities related to a dynamical system associated with the graph. It will be particularly challenging to consider directed graphs, and to impose that the algorithm is anonymous, i.e., that it does not makes use of labels identifying the different agents associated with vertices.

3.4. Collaborative and distributed network control

This research line deals with the problem of designing controllers with a limited use of the network information (i.e. with restricted feedback), and with the aim to reach a pre-specified global behavior. This is in contrast to centralized controllers that use the whole system information and compute the control law at some central node. Collaborative control has already been explored in the team in connection with the underwater robot fleet, and to some extent with the source seeking problem. It remains however a certain number of challenging problems that the team wishes to address:

- Design of control with limited information, able to lead to desired global behaviors. Here the graph structure is imposed by the problem, and we aim to design the "best" possible control under such a graph constraint [0]. The team would like to explore further this research line, targeting a better understanding of possible metrics to be used as a target for optimal control design. In particular, and in connection with the traffic application, the long-standing open problem of ramp metering control under minimum information will be addressed.
- Clustering control for large networks. For large and complex systems composed of several sub-networks, feedback design is usually treated at the sub-network level, and most of the times without taking into account natural interconnections between sub-networks. The team is exploring new control strategies, exploiting the emergent behaviors resulting from new interconnections between the network components. This requires first to build network models operating in aggregated clusters, and then to re-formulate problems where the control can be designed using the cluster boundaries rather than individual control loops inside of each network. Examples can be found in the transportation application domain, where a significant challenge will be to obtain dynamic partitioning and clustering of heterogeneous networks in homogeneous sub-networks, and then to control the perimeter flows of the clusters to optimize the network operation. This topic is at the core of the Advanced ERC project Scale-FreeBack.

[0] Such a problem has been previously addressed in some specific applications, particularly robot fleets, and only few recent theoretical works have initiated a more systematic system-theoretic study of sparsity-constrained system realization theory and of sparsity-constrained feedback control.

3.5. Transportation networks

This is currently the main application domain of the NECS team. Several interesting problems in this area capture many of the generic networks problems identified before (e.g., decentralized/collaborative traffic optimal control, density balancing using consensus concepts, data fusion, distributed estimation, etc.). Several specific actions have been continued/launched to this purpose: improvement and finalization of the Grenoble Traffic Lab (GTL), EU projects (SPEEDD, ERC-AdG Scale-FreeBack). Further research goals are envisioned, such as:

- Modeling of large scale traffic systems. We aim at reducing the complexity of traffic systems modeling by engaging novel modeling techniques that make use of clustering for traffic networks while relying on its specific characteristics. Traffic networks will be aggregate into clusters and the main traffic quantities will be extrapolated by making use of this aggregation. Moreover, we are developing an extension of the Grenoble Traffic Lab (GTL) for downtown Grenoble which will make use of GPS and probe data to collect traffic data in the city center.
- Modeling and control of intelligent transportation systems. We aim at developing a complete micro-macro modeling approach to describe and model the new traffic dynamics that is developing thanks to mixed (simple, connected and automated) vehicles in the roads. This will require cutting edge mathematical theory and field experiments.

POEMS Project-Team

3. Research Program

3.1. Expertises

The activity of the team is oriented towards the design, the analysis and the numerical approximation of mathematical models for all types of problems involving wave propagation phenomena, in mechanics, physics and engineering sciences. Let us briefly describe our core business and current expertise, in order to clarify the new challenges that we want to address in the short and long terms.

Typically, our works are based on *boundary value problems* established by physicists to model the propagation of waves in various situations. The basic ingredient is a partial differential equation of the hyperbolic type, whose prototype is the scalar wave equation, or the Helmholtz equation if time-periodic solutions are considered. More generally, we systematically consider both the transient problem, in the time domain, and the time-harmonic problem, in the frequency domain. Let us mention that, even if different waves share a lot of common properties, the transition from the scalar acoustic equation to the vectorial electromagnetism and elastodynamics systems raises a lot of mathematical and numerical difficulties, and requires a specific expertise.

A notable particularity of the problems that we consider is that they are generally set in *unbounded domains*: for instance, for radar applications, it is necessary to simulate the interaction of the electromagnetic waves with the airplane only, without any complex environment perturbing the wave phenomena. This raises an intense research activity, both from a theoretical and a numerical point of view. There exist several approaches which all consist in rewriting the problem (or an approximation of it) in a bounded domain, the new formulation being well-suited for classical mathematical and numerical techniques.

One class of methods consists in applying an appropriate condition on some boundary enclosing the zone of interest. In the frequency domain, one can use a non-local transparent condition, which can be expressed by a convolution with a Green function like in integral equation techniques, or by a modal decomposition when a separation of variables is applicable. But for explicit schemes in the time domain, local radiation conditions at a finite distance are generally preferred (constructed as local approximations at various orders of the exact non-local condition). A second class of methods consists in surrounding the computational domain by so called *Perfectly Matched absorbing Layers* (PML), which are very popular because they are easy to implement. POEMS members have provided several contributions to these two classes of methods for more than twenty-five years. Among them, on can mention the understanding of the instability of PMLs in anisotropic media and in dispersive media, the derivation of transparent boundary conditions in periodic media or the improvement of Fast Multipole techniques for elastodynamic integral equations.

In addition to more classical domains of applied mathematics that we are led to use (variational analysis and functional analysis, interpolation and approximation theory, linear algebra of large systems, etc...), we have acquired a deep expertise in *spectral theory*. Indeed, the analysis of wave phenomena is intimately linked to the study of some associated spectral problems. Acoustic resonance frequencies of a cavity correspond to the eigenvalues of a selfadjoint Laplacian operator, modal solutions in a waveguide correspond to a spectral problem set in the cross section. In these two examples, if the cavity or the cross-section is unbounded, a part of the spectrum is a continuum. Again, POEMS has produced several contributions in this field. In particular, a large number of significant results have been obtained for the existence or non-existence of guided modes in open waveguides and of trapped modes in infinite domains.

To end this far from exhaustive presentation of our main expertise domains, let us mention the *asymptotic techniques* with respect to some small scale appearing in the model: it can be the wavelength compared to the size of the scatterer, or on the contrary, the scale of the scatterer compared to the wavelength, it can be the scale of some microstructure in a composite material or the width of a thin layer or a thin tube. In each case, the objective, in order to avoid the use of costly meshes, is to derive effective simplified models. Our

specificity here is that we can combine skills in physics, mathematics and numerics: in particular, we take care of the mathematical properties of the effective model, which are used to ensure the robustness of the numerical method, and also to derive error estimates with respect to the small parameter. There has been a lot of contributions of POEMS to this topic, going from the modeling of electromagnetic coatings to the justification of models for piezoelectric sensors. Let us mention that effective models for small scatterers and thin coatings have been used to improve imaging techniques that we are developing (topological gradient, time reversal or sampling techniques).

3.2. New domains

In order to consider more and more challenging problems (involving non-deterministic, large-scale and more realistic models), we decided recently to enlarge our domain of expertise in three directions.

Firstly, we want to reinforce our activity on *efficient solvers for large-scale wave propagation problems*. Since its inception, POEMS has frequently contributed to the development and the analysis of numerical methods that permit the fast solution of large-scale problems, such as high-order finite element methods, boundary elements methods and domain decomposition methods. Nevertheless, implementing these methods in parallel programming environments and dealing with large-scale benchmarks have generally not been done by the team. We want to continue our activities on these methods and, in a more comprehensive approach, we will incorporate modern algebraic strategies and high-performance computing (HPC) aspects in our methodology. In collaboration with academic or industrial partners, we would like to address industrial-scale benchmarks to assess the performance of our approaches. We believe that taking all these aspects into consideration will allow us to design more efficient wave-specific computational tools for large-scale simulations.

Secondly, up to now, *probabilistic methods* were outside the expertise of POEMS team, restricting us to deterministic approaches for wave propagation problems. We however firmly believe in the importance and usefulness of addressing uncertainty and randomness inherent to many propagation phenomena. Randomness may occur in the description of complex propagation media (for example in the modeling of ultrasound waves in concrete for the simulation of non-destructive testing experiments) or of data uncertainties. To quantify the effect of such uncertainties on the design, behavior, performance or reliability of many systems is then a natural goal in diverse fields of application.

Thirdly and lastly, we wish to develop and strengthen collaborations allowing a *closer interaction between our mathematical, modeling and computing activities and physical experiments*, where the latter may either provide reality checks on existing models or strongly affect the choice of modeling assumptions. Within our typical domain of activities, we can mention three areas for which such considerations are highly relevant. One is musical acoustics, where POEMS has made several well-recognized contributions dealing with the simulation of musical instruments. Another area is inverse problems, whose very purpose is to extract useful information from actual measurements with the help of (propagation) models. This is a core of our partnership with CEA on ultrasonic Non Destructive Testing. A third area is the modelling of effective (acoustic or electromagnetic) metamaterials, where predictions based on homogenized models have to be confirmed by experiments.

QUANTIC Project-Team

3. Research Program

3.1. Hardware-efficient quantum information processing

In this scientific program, we will explore various theoretical and experimental issues concerning protection and manipulation of quantum information. Indeed, the next, critical stage in the development of Quantum Information Processing (QIP) is most certainly the active quantum error correction (QEC). Through this stage one designs, possibly using many physical qubits, an encoded logical qubit which is protected against major decoherence channels and hence admits a significantly longer effective coherence time than a physical qubit. Reliable (fault-tolerant) computation with protected logical qubits usually comes at the expense of a significant overhead in the hardware (up to thousands of physical qubits per logical qubit). Each of the involved physical qubits still needs to satisfy the best achievable properties (coherence times, coupling strengths and tunability). More remarkably, one needs to avoid undesired interactions between various subsystems. This is going to be a major difficulty for qubits on a single chip.

The usual approach for the realization of QEC is to use many qubits to obtain a larger Hilbert space of the qubit register [88], [92]. By redundantly encoding quantum information in this Hilbert space of larger dimension one make the QEC tractable: different error channels lead to distinguishable error syndromes. There are two major drawbacks in using multi-qubit registers. The first, fundamental, drawback is that with each added physical qubit, several new decoherence channels are added. Because of the exponential increase of the Hilbert's space dimension versus the linear increase in the number of decay channels, using enough qubits, one is able to eventually protect quantum information against decoherence. However, multiplying the number of possible errors, this requires measuring more error syndromes. Note furthermore that, in general, some of these new decoherence channels can lead to correlated action on many qubits and this needs to be taken into account with extra care: in particular, such kind of non-local error channels are problematic for surface codes. The second, more practical, drawback is that it is still extremely challenging to build a register of more than on the order of 10 qubits where each of the qubits is required to satisfy near the best achieved properties: these properties include the coherence time, the coupling strengths and the tunability. Indeed, building such a register is not merely only a fabrication task but rather, one requirers to look for architectures such that, each individual qubit can be addressed and controlled independently from the others. One is also required to make sure that all the noise channels are well-controlled and uncorrelated for the QEC to be effective.

We have recently introduced a new paradigm for encoding and protecting quantum information in a quantum harmonic oscillator (e.g. a high-Q mode of a 3D superconducting cavity) instead of a multi-qubit register [64]. The infinite dimensional Hilbert space of such a system can be used to redundantly encode quantum information. The power of this idea lies in the fact that the dominant decoherence channel in a cavity is photon damping, and no more decay channels are added if we increase the number of photons we insert in the cavity. Hence, only a single error syndrome needs to be measured to identify if an error has occurred or not. Indeed, we are convinced that most early proposals on continuous variable QIP [61], [55] could be revisited taking into account the design flexibilities of Quantum Superconducting Circuits (QSC) and the new coupling regimes that are provided by these systems. In particular, we have illustrated that coupling a qubit to the cavity mode in the strong dispersive regime provides an important controllability over the Hilbert space of the cavity mode [63]. Through a recent experimental work [97], we benefit from this controllability to prepare superpositions of quasi-orthogonal coherent states, also known as Schrödinger cat states.

In this Scheme, the logical qubit is encoded in a four-component Schrödinger cat state. Continuous quantum non-demolition (QND) monitoring of a single physical observable, consisting of photon number parity, enables then the tractability of single photon jumps. We obtain therefore a first-order quantum error correcting code using only a single high-Q cavity mode (for the storage of quantum information), a single qubit (providing the non-linearity needed for controllability) and a single low-Q cavity mode (for reading out the error syndrome).

An earlier experiment on such QND photon-number parity measurements [93] has recently led to a first experimental realization of a full quantum error correcting code improving the coherence time of quantum information [8]. As shown in Figure 1 , this leads to a significant hardware economy for realization of a protected logical qubit. Our goal here is to push these ideas towards a reliable and hardware-efficient paradigm for universal quantum computation.

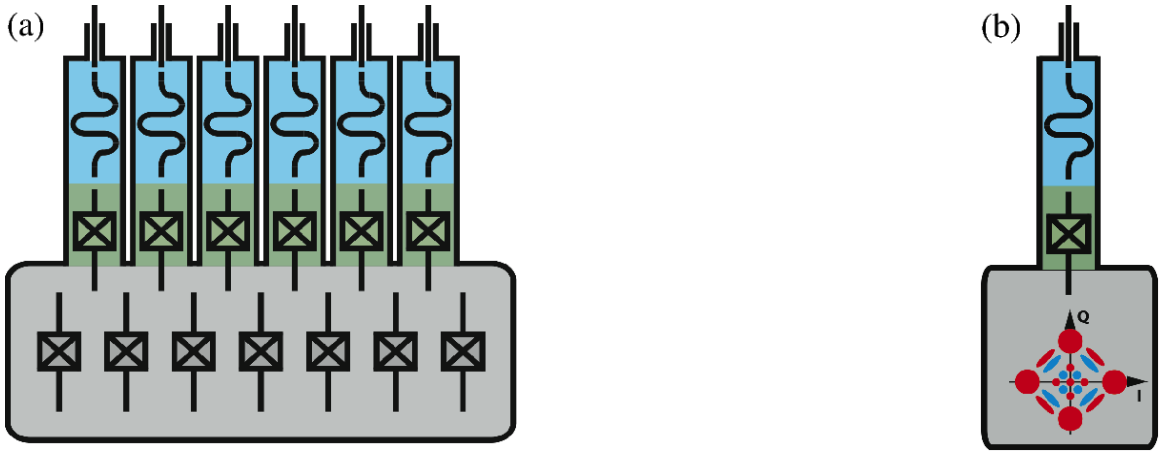


Figure 1. (a) A protected logical qubit consisting of a register of many qubits: here, we see a possible architecture for the Steane code [92] consisting of 7 qubits requiring the measurement of 6 error syndromes. In this sketch, 7 transmon qubits in a high-Q resonator and the measurement of the 6 error syndromes is ensured through 6 additional ancillary qubits with the possibility of individual readout of the ancillary qubits via independent low-Q resonators. (b) Minimal architecture for a protected logical qubit, adapted to circuit quantum electrodynamics experiments. Quantum information is encoded in a Schrödinger cat state of a single high-Q resonator mode and a single error syndrome is measured, using a single ancillary transmon qubit and the associated readout low-Q resonator.

3.2. Reservoir (dissipation) engineering and autonomous stabilization of quantum systems

Being at the heart of any QEC protocol, the concept of feedback is central for the protection of quantum information, enabling many-qubit quantum computation or long-distance quantum communication. However, such a closed-loop control which requires a real-time and continuous measurement of the quantum system has been for long considered as counter-intuitive or even impossible. This thought was mainly caused by properties of quantum measurements: any measurement implies an instantaneous strong perturbation to the system's state. The concept of *quantum non-demolition* (QND) measurement has played a crucial role in understanding and resolving this difficulty [37]. In the context of cavity quantum electro-dynamics (cavity QED) with Rydberg atoms [57], a first experiment on continuous QND measurements of the number of microwave photons was performed by the group at Laboratoire Kastler-Brossel (ENS) [56]. Later on, this ability of performing continuous measurements allowed the same group to realize the first continuous quantum feedback protocol stabilizing highly non-classical states of the microwave field in the cavity, the so-called photon number states [10] (this ground-breaking work was mentioned in the Nobel prize attributed to Serge Haroche). The QUANTIC team contributed to the theoretical work behind this experiment [47], [29], [91] [1]. These contributions include the development and optimization of the quantum filters taking into account the quantum measurement back-action and various measurement noises and uncertainties, the development of a feedback law based on control Lyapunov techniques, and the compensation of the feedback delay.

In the context of circuit quantum electrodynamics (circuit QED) [45], recent advances in quantum-limited amplifiers [81], [95] have opened doors to high-fidelity non-demolition measurements and real-time feedback for superconducting qubits [58]. This ability to perform high-fidelity non-demolition measurements of a quantum signal has very recently led to quantum feedback experiments with quantum superconducting circuits [95], [80], [39]. Here again, the QUANTIC team has participated to one of the first experiments in the field where the control objective is to track a dynamical trajectory of a single qubit rather than stabilizing a stationary state. Such quantum trajectory tracking could be further explored to achieve metrological goals such as the stabilization of the amplitude of a microwave drive [71].

While all this progress has led to a strong optimism about the possibility to perform active protection of quantum information against decoherence, the rather short dynamical time scales of these systems limit, to a great amount, the complexity of the feedback strategies that could be employed. Indeed, in such measurement-based feedback protocols, the time-consuming data acquisition and post-treatment of the output signal leads to an important latency in the feedback procedure.

The reservoir (dissipation) engineering [78] and the closely related coherent feedback [69] are considered as alternative approaches circumventing the necessity of a real-time data acquisition, signal processing and feedback calculations. In the context of quantum information, the decoherence, caused by the coupling of a system to uncontrolled external degrees of freedom, is generally considered as the main obstacle to synthesize quantum states and to observe quantum effects. Paradoxically, it is possible to intentionally engineer a particular coupling to a reservoir in the aim of maintaining the coherence of some particular quantum states. In a general viewpoint, these approaches could be understood in the following manner: by coupling the quantum system to be stabilized to a strongly dissipative ancillary quantum system, one evacuates the entropy of the main system through the dissipation of the ancillary one. By building the feedback loop into the Hamiltonian, this type of autonomous feedback obviates the need for a complicated external control loop to correct errors. On the experimental side, such autonomous feedback techniques have been used for qubit reset [54], single-qubit state stabilization [73], and the creation [32] and stabilization [62], [68], [87] of states of multipartite quantum systems.

Such reservoir engineering techniques could be widely revisited exploring the flexibility in the Hamiltonian design for QSC. We have recently developed theoretical proposals leading to extremely efficient, and simple to implement, stabilization schemes for systems consisting of a single, two or three qubits [54], [66], [43], [46]. The experimental results based on these protocols have illustrated the efficiency of the approach [54], [87]. Through these experiments, we exploit the strong dispersive interaction [85] between superconducting qubits and a single low-Q cavity mode playing the role of a dissipative reservoir. Applying continuous-wave (cw) microwave drives with well-chosen fixed frequencies, amplitudes, and phases, we engineer an effective interaction Hamiltonian which evacuates the entropy of the system interacting with a noisy environment: by driving the qubits and cavity with continuous-wave drives, we induce an autonomous feedback loop which corrects the state of the qubits every time it decays out of the desired target state. The schemes are robust against small variations of the control parameters (drives amplitudes and phase) and require only some basic calibration. Finally, by avoiding resonant interactions between the qubits and the low-Q cavity mode, the qubits remain protected against the Purcell effect, which would reduce the coherence times. We have also investigated both theoretically and experimentally the autonomous stabilization of non-classical states (such as Schrodinger cat states and Fock states) of microwave field confined in a high-Q cavity mode [83], [59][7], [6].

3.3. System theory for quantum information processing

In parallel and in strong interactions with the above experimental goals, we develop systematic mathematical methods for dynamical analysis, control and estimation of composite and open quantum systems. These systems are built with several quantum subsystems whose irreversible dynamics results from measurements and/or decoherence. A special attention is given to spin/spring systems made with qubits and harmonic oscillators. These developments are done in the spirit of our recent contributions [82], [29], [90], [84], [91][9], [1] resulting from collaborations with the cavity quantum electrodynamics group of Laboratoire Kastler Brossel.

3.3.1. Stabilization by measurement-based feedback

The protection of quantum information via efficient QEC is a combination of (i) tailored dynamics of a quantum system in order to protect an informational qubit from certain decoherence channels, and (ii) controlled reaction to measurements that efficiently detect and correct the dominating disturbances that are not rejected by the tailored quantum dynamics.

In such feedback scheme, the system and its measurement are quantum objects whereas the controller and the control input are classical. The stabilizing control law is based on the past values of the measurement outcomes. During our work on the LKB photon box, we have developed, for single input systems subject to quantum non-demolition measurement, a systematic stabilization method [1]: it is based on a discrete-time formulation of the dynamics, on the construction of a strict control Lyapunov function and on an explicit compensation of the feedback-loop delay. Keeping the QND measurement assumptions, extensions of such stabilization schemes will be investigated in the following directions: finite set of values for the control input with application to the convergence analysis of the atomic feedback scheme experimentally tested in [98]; multi-input case where the construction by inversion of a Metzler matrix of the strict Lyapunov function is not straightforward; continuous-time systems governed by diffusive master equations; stabilization towards a set of density operators included in a target subspace; adaptive measurement by feedback to accelerate the convergence towards a stationary state as experimentally tested in [76]. Without the QND measurement assumptions, we will also address the stabilization of non-stationary states and trajectory tracking, with applications to systems similar to those considered in [58], [39].

3.3.2. Filtering, quantum state and parameter estimations

The performance of every feedback controller crucially depends on its online estimation of the current situation. This becomes even more important for quantum systems, where full state measurements are physically impossible. Therefore the ultimate performance of feedback correction depends on fast, efficient and optimally accurate state and parameter estimations.

A quantum filter takes into account imperfection and decoherence and provides the quantum state at time $t \geq 0$ from an initial value at $t = 0$ and the measurement outcomes between 0 and t. Quantum filtering goes back to the work of Belavkin [33] and is related to quantum trajectories [41], [44]. A modern and mathematical exposure of the diffusive models is given in [31]. In [99] a first convergence analysis of diffusive filters is proposed. Nevertheless the convergence characterization and estimation of convergence rate remain open and difficult problems. For discrete time filters, a general stability result based on fidelity is proven in [82], [90]. This stability result is extended to a large class of continuous-time filters in [30]. Further efforts are required to characterize asymptotic and exponential stability. Estimations of convergence rates are available only for quantum non-demolition measurements [34]. Parameter estimations based on measurement data of quantum trajectories can be formulated within such quantum filtering framework [49], [74].

We will continue to investigate stability and convergence of quantum filtering. We will also exploit our fidelity-based stability result to justify maximum likelihood estimation and to propose, for open quantum system, parameter estimation algorithms inspired of existing estimation algorithms for classical systems. We will also investigate a more specific quantum approach: it is noticed in [38] that post-selection statistics and "past quantum" state analysis [50] enhance sensitivity to parameters and could be interesting towards increasing the precision of an estimation.

3.3.3. Stabilization by interconnections

In such stabilization schemes, the controller is also a quantum object: it is coupled to the system of interest and is subject to decoherence and thus admits an irreversible evolution. These stabilization schemes are closely related to reservoir engineering and coherent feedback [78], [69]. The closed-loop system is then a composite system built with the original system and its controller. In fact, and given our particular recent expertise in this domain [9] [87], [54], this subsection is dedicated to further developing such stabilization techniques, both experimentally and theoretically.

The main analysis issues are to prove the closed-loop convergence and to estimate the convergence rates. Since these systems are governed by Lindblad differential equations (continuous-time case) or Kraus maps (discrete-time case), their stability is automatically guaranteed: such dynamics are contractions for a large set of metrics (see [77]). Convergence and asymptotic stability is less well understood. In particular most of the convergence results consider the case where the target steady-state is a density operator of maximum rank (see, e.g., [28][chapter 4, section 6]). When the goal steady-state is not full rank very few convergence results are available.

We will focus on this geometric situation where the goal steady-state is on the boundary of the cone of positive Hermitian operators of finite trace. A specific attention will be given to adapt standard tools (Lyapunov function, passivity, contraction and Lasalle's invariance principle) for infinite dimensional systems to spin/spring structures inspired of [9], [7] [87], [54] and their associated Fokker-Planck equations for the Wigner functions.

We will also explore the Heisenberg point of view in connection with recent results of the Inria project-team MAXPLUS (algorithms and applications of algebras of max-plus type) relative to Perron-Frobenius theory [53], [52]. We will start with [86] and [79] where, based on a theorem due to Birkhoff [35], dual Lindblad equations and dual Kraus maps governing the Heisenberg evolution of any operator are shown to be contractions on the cone of Hermitian operators equipped with Hilbert's projective metric. As the Heisenberg picture is characterized by convergence of all operators to a multiple of the identity, it might provide a mean to circumvent the rank issues. We hope that such contraction tools will be especially well adapted to analyzing quantum systems composed of multiple components, motivated by the facts that the same geometry describes the contraction of classical systems undergoing synchronizing interactions [94] and by our recent generalized extension of the latter synchronizing interactions to quantum systems [70].

Besides these analysis tasks, the major challenge in stabilization by interconnections is to provide systematic methods for the design, from typical building blocks, of control systems that stabilize a specific quantum goal (state, set of states, operation) when coupled to the target system. While constructions exist for so-called linear quantum systems [75], this does not cover the states that are more interesting for quantum applications. Various strategies have been proposed that concatenate iterative control steps for open-loop steering [96], [67] with experimental limitations. The characterization of Kraus maps to stabilize any types of states has also been established [36], but without considering experimental implementations. A viable stabilization by interaction has to combine the capabilities of these various approaches, and this is a missing piece that we want to address.

3.3.3.1. Perturbation methods

With this subsection we turn towards more fundamental developments that are necessary in order to address the complexity of quantum networks with efficient reduction techniques. This should yield both efficient mathematical methods, as well as insights towards unravelling dominant physical phenomena/mechanisms in multipartite quantum dynamical systems.

In the Schrödinger point of view, the dynamics of open quantum systems are governed by master equations, either deterministic or stochastic [57], [51]. Dynamical models of composite systems are based on tensor products of Hilbert spaces and operators attached to the constitutive subsystems. Generally, a hierarchy of different timescales is present. Perturbation techniques can be very useful to construct reliable models adapted to the timescale of interest.

To eliminate high frequency oscillations possibly induced by quasi-resonant classical drives, averaging techniques are used (rotating wave approximation). These techniques are well established for closed systems without any dissipation nor irreversible effect due to measurement or decoherence. We will consider in a first step the adaptation of these averaging techniques to deterministic Lindblad master equations governing the quantum state, i.e. the system density operator. Emphasis will be put on first order and higher order corrections based on non-commutative computations with the different operators appearing in the Lindblad equations. Higher order terms could be of some interest for the protected logical qubit of figure 1 b. In future steps, we intend to explore the possibility to explicitly exploit averaging or singular perturbation properties in the design of coherent quantum feedback systems; this should be an open-systems counterpart of works like [65].

To eliminate subsystems subject to fast convergence induced by decoherence, singular perturbation techniques can be used. They provide reduced models of smaller dimension via the adiabatic elimination of the rapidly converging subsystems. The derivation of the slow dynamics is far from being obvious (see, e.g., the computations of page 142 in [40] for the adiabatic elimination of low-Q cavity). Conversely to the classical composite systems where we have to eliminate one component in a Cartesian product, we here have to eliminate one component in a tensor product. We will adapt geometric singular perturbations [48] and invariant manifold techniques [42] to such tensor product computations to derive reduced slow approximations of any order. Such adaptations will be very useful in the context of quantum Zeno dynamics to obtain approximations of the slow dynamics on the decoherence-free subspace corresponding to the slow attractive manifold.

Perturbation methods are also precious to analyze convergence rates. Deriving the spectrum attached to the Lindblad differential equation is not obvious. We will focus on the situation where the decoherence terms of the form $L\rho L^\dagger - (L^\dagger L\rho + \rho L^\dagger L)/2$ are small compared to the conservative terms $-i[H/\hbar, \rho]$. The difficulty to overcome here is the degeneracy of the unperturbed spectrum attached to the conservative evolution $\frac{d}{dt}\rho = -i[H/\hbar, \rho]$. The degree of degeneracy of the zero eigenvalue always exceeds the dimension of the Hilbert space. Adaptations of usual perturbation techniques [60] will be investigated. They will provide estimates of convergence rates for slightly open quantum systems. We expect that such estimates will help to understand the dependence on the experimental parameters of the convergence rates observed in [54], [87], [66].

As particular outcomes for the other subsections, we expect that these developments towards simpler dominant dynamics will guide the search for optimal control strategies, both in open-loop microwave networks and in autonomous stabilization schemes such as reservoir engineering. It will further help to efficiently compute explicit convergence rates and quantitative performances for all the intended experiments.

RANDOPT Project-Team

3. Research Program

3.1. Introduction

The lines of research we intend to pursue is organized along four axis namely developing novel theoretical framework, developing novel algorithms, setting novel standards in scientific experimentation and benchmarking and applications.

3.2. Developing Novel Theoretical Frameworks for Analyzing and Designing Adaptive Stochastic Algorithms

Stochastic black-box algorithms typically optimize **non-convex, non-smooth functions**. This is possible because the algorithms rely on weak mathematical properties of the underlying functions: the algorithms do not use the derivatives—hence the function does not need to be differentiable—and, additionally, often do not use the exact function value but instead how the objective function ranks candidate solutions (such methods are sometimes called function-value-free).(To illustrate a comparison-based update, consider an algorithm that samples λ (with λ an even integer) candidate solutions from a multivariate normal distribution. Let $x_1, ..., x_\lambda$ in $\mathbb{R}^n$ denote those λ candidate solutions at a given iteration. The solutions are evaluated on the function f to be minimized and ranked from the best to the worse:

$$f(x_{1:\lambda}) \leq ... \leq f(x_{\lambda:\lambda}) \ .$$

In the previous equation $i:\lambda$ denotes the index of the sampled solution associated to the i-th best solution. The new mean of the Gaussian vector from which new solutions will be sampled at the next iteration can be updated as

$$m \longleftarrow \frac{1}{\lambda} \sum_{i=1}^{\lambda/2} x_{i:\lambda} \ .$$

The previous update moves the mean towards the $\lambda/2$ best solutions. Yet the update is only based on the ranking of the candidate solutions such that the update is the same if f is optimized or $g \circ f$ where $g : \mathrm{Im}(f) \to \mathbb{R}$ is strictly increasing. Consequently, such algorithms are invariant with respect to strictly increasing transformations of the objective function. This entails that they are robust and their performances generalize well.)

Additionally, adaptive stochastic optimization algorithms typically have a **complex state space** which encodes the parameters of a probability distribution (e.g. mean and covariance matrix of a Gaussian vector) and other state vectors. This state-space is a **manifold**. While the algorithms are Markov chains, the complexity of the state-space makes that **standard Markov chain theory tools do not directly apply**. The same holds with tools stemming from stochastic approximation theory or Ordinary Differential Equation (ODE) theory where it is usually assumed that the underlying ODE (obtained by proper averaging and limit for learning rate to zero) has its critical points inside the search space. In contrast, in the cases we are interested in, the **critical points of the ODEs are at the boundary of the domain**.

Last, since we aim at developing theory that on the one hand allows to analyze the main properties of state-of-the-art methods and on the other hand is useful for algorithm design, we need to be careful not to use simplifications that would allow a proof to be done but would not capture the important properties of the algorithms. With that respect one tricky point is to develop **theory that accounts for invariance properties**.

To face those specific challenges, we need to develop novel theoretical frameworks exploiting invariance properties and accounting for peculiar state-spaces. Those frameworks should allow researchers to analyze one of the core properties of adaptive stochastic methods, namely **linear convergence** on the widest possible class of functions.

We are planning to approach the question of linear convergence from three different complementary angles, using three different frameworks:

- the Markov chain framework where the convergence derives from the analysis of the stability of a normalized Markov chain existing on scaling-invariant functions for translation and scale-invariant algorithms [18]. This framework allows for a fine analysis where the exact convergence rate can be given as an implicit function of the invariant measure of the normalized Markov chain. Yet it requires the objective function to be scaling-invariant. The stability analysis can be particularly tricky as the Markov chain that needs to be studied writes as $\Phi_{t+1} = F(\Phi_t, W_{t+1})$ where $\{W_t : t > 0\}$ are independent identically distributed and F is typically discontinuous because the algorithms studied are comparison-based. This implies that practical tools for analyzing a standard property like irreducibility, that rely on investigating the stability of underlying deterministic control models [34], cannot be used. Additionally, the construction of a drift to prove ergodicity is particularly delicate when the state space includes a (normalized) covariance matrix as it is the case for analyzing the CMA-ES algorithm.
- The stochastic approximation or ODE framework. Those are standard techniques to prove the convergence of stochastic algorithms when an algorithm can be expressed as a stochastic approximation of the solution of a mean field ODE [20], [21], [32]. What is specific and induces difficulties for the algorithms we aim at analyzing is the **non-standard state-space** since the ODE variables correspond to the state-variables of the algorithm (e.g. $\mathbb{R}^n \times \mathbb{R}_{>0}$ for step-size adaptive algorithms, $\mathbb{R}^n \times \mathbb{R}_{>0} \times S^n_{++}$ where S^n_{++} denotes the set of positive definite matrices if a covariance matrix is additionally adapted). Consequently, the ODE can have many critical points at the boundary of its definition domain (e.g. all points corresponding to $\sigma_t = 0$ are critical points of the ODE) which is not typical. Also we aim at proving **linear convergence**, for that it is crucial that the learning rate does not decrease to zero which is non-standard in ODE method.
- The direct framework where we construct a global Lyapunov function for the original algorithm from which we deduce bounds on the hitting time to reach an ϵ-ball of the optimum. For this framework as for the ODE framework, we expect that the class of functions where we can prove linear convergence are composite of $g \circ f$ where f is differentiable and $g : \mathrm{Im}(f) \to \mathbb{R}$ is strictly increasing and that we can show convergence to a local minimum.

We expect those frameworks to be complementary in the sense that the assumptions required are different. Typically, the ODE framework should allow for proofs under the assumptions that learning rates are small enough while it is not needed for the Markov chain framework. Hence this latter framework captures better the real dynamics of the algorithm, yet under the assumption of scaling-invariance of the objective functions. Also, we expect some overlap in terms of function classes that can be studied by the different frameworks (typically convex-quadratic functions should be encompassed in the three frameworks). By studying the different frameworks in parallel, we expect to gain synergies and possibly understand what is the most promising approach for solving the holy grail question of the linear convergence of CMA-ES. We foresee for instance that similar approaches like the use of Foster-Lyapunov drift conditions are needed in all the frameworks and that intuition can be gained on how to establish the conditions from one framework to another one.

3.3. Algorithmic developments

We are planning on developing algorithms in the subdomains with strong practical demand for better methods of constrained, multiobjective, large-scale and expensive optimization.

Many of the algorithm developments, we propose, rely on the CMA-ES method. While this seems to restrict our possibilities, we want to emphasize that CMA-ES became a *family of methods* over the years that nowadays include various techniques and developments from the literature to handle non-standard optimization problems (noisy, large-scale, ...). The core idea of all CMA-ES variants—namely the mechanism to adapt a Gaussian distribution—has furthermore been shown to derive naturally from first principles with only minimal assumptions in the context of derivative-free black-box stochastic optimization [35], [25]. This is a strong justification for relying on the CMA-ES premises while new developments naturally include new techniques typically borrowed from other fields. While CMA-ES is now a full family of methods, for visibility reasons, we continue to refer often to "the CMA-ES algorithm".

3.3.1. Constrained optimization

Many (real-world) optimization problems have constraints related to technical feasibility, cost, etc. Constraints are classically handled in the black-box setting either via rejection of solutions violating the constraints—which can be quite costly and even lead to quasi-infinite loops—or by penalization with respect to the distance to the feasible domain (if this information can be extracted) or with respect to the constraint function value [22]. However, the penalization coefficient is a sensitive parameter that needs to be adapted in order to achieve a robust and general method [23]. Yet, **the question of how to handle properly constraints is largely unsolved**. The latest constraints handling for CMA-ES is an ad-hoc technique driven by many heuristics [23]. Also, it is particularly only recently that it was pointed out that **linear convergence properties should be preserved** when addressing constraint problems [16].

Promising approaches though, rely on using augmented Lagrangians [16], [17]. The augmented Lagrangian, here, is the objective function optimized by the algorithm. Yet, it depends on coefficients that are adapted online. The adaptation of those coefficients is the difficult part: the algorithm should be stable and the adaptation efficient. We believe that the theoretical frameworks developed (particularly the Markov chain framework) will be useful to understand how to design the adaptation mechanisms. Additionally, the question of invariance will also be at the core of the design of the methods: augmented Lagrangian approaches break the invariance to monotonic transformation of the objective functions, yet understanding the maximal invariance that can be achieved seems to be an important step towards understanding what adaptation rules should satisfy.

3.3.2. Large-scale Optimization

In the large-scale setting, we are interested to optimize problems with the order of 10^3 to 10^4 variables. For one to two orders of magnitude more variables, we will talk about a "very large-scale" setting.

In this context, algorithms with a quadratic scaling (internal and in terms of number of function evaluations needed to optimize the problem) cannot be afforded. In CMA-ES-type algorithms, we typically need to restrict the model of the covariance matrix to have only a linear number of parameters to learn such that the algorithms scale linearly in terms of internal complexity, memory and number of function evaluations to solve the problem. The main challenge is thus to have rich enough models for which we can efficiently design proper adaptation mechanisms. Some first large-scale variants of CMA-ES have been derived. They include the online adaptation of the complexity of the model [15], [14]. Yet so far they fail to optimize functions whose Hessian matrix has some small eigenvalues (say around 10^{-4}) some eigenvalues equal to 1 and some very large eigenvalue (say around 10^4), that is functions whose level sets have short and long axis.

Another direction, we want to pursue, is exploring the use of large-scale variants of CMA-ES to solve reinforcement learning problems [36].

Last, we are interested to investigate the very-large-scale setting. One approach consists in doing optimization in subspaces. This entails the efficient identification of relevant spaces and the restriction of the optimization to those subspaces.

3.3.3. Multiobjective Optimization

Multiobjective optimization, i.e., the simultaneous optimization of multiple objective functions, differs from single-objective optimization in particular in its optimization goal. Instead of aiming at converging to the

solution with the best possible function value, in multiobjective optimization, a set of solutions [0] is sought. This set, called Pareto-set, contains all trade-off solutions in the sense of Pareto-optimality—no solution exists that is better in *all* objectives than a Pareto-optimal one. Because converging towards a set differs from converging to a single solution, it is no surprise that we might lose many good convergence properties if we directly apply search operators from single-objective methods. However, this is what has typically been done so far in the literature. Indeed, most of the research in stochastic algorithms for multiobjective optimization focused instead on the so called selection part, that decides which solutions should be kept during the optimization—a question that can be considered as solved for many years in the case of single-objective stochastic adaptive methods.

We therefore aim at rethinking search operators and adaptive mechanisms to improve existing methods. We expect that we can obtain orders of magnitude better convergence rates for certain problem types if we choose the right search operators. We typically see two angles of attack: On the one hand, we will study methods based on scalarizing functions that transform the multiobjective problem into a set of single-objective problems. Those single-objective problems can then be solved with state-of-the-art single-objective algorithms. Classical methods for multiobjective optimization fall into this category, but they all solve multiple single-objective problems subsequently (from scratch) instead of dynamically changing the scalarizing function during the search. On the other hand, we will improve on currently available population-based methods such as the first multiobjective versions of the CMA-ES. Here, research is needed on an even more fundamental level such as trying to understand success probabilities observed during an optimization run or how we can introduce non-elitist selection (the state of the art in single-objective stochastic adaptive algorithms) to increase robustness regarding noisy evaluations or multi-modality. The challenge here, compared to single-objective algorithms, is that the quality of a solution is not anymore independent from other sampled solutions, but can potentially depend on all known solutions (in the case of three or more objective functions), resulting in a more noisy evaluation as the relatively simple function-value-based ranking within single-objective optimizers.

3.3.4. Expensive Optimization

In the so-called expensive optimization scenario, a single function evaluation might take several minutes or even hours in a practical setting. Hence, the available budget in terms of number of function evaluation calls to find a solution is very limited in practice. To tackle such expensive optimization problems, it is needed to exploit the first few function evaluations in the best way. To this end, typical methods couple the learning of a surrogate (or meta-model) of the expensive objective function with traditional optimization algorithms.

In the context of expensive optimization and CMA-ES, which usually shows its full potential when the number n of variables is not too small (say larger than 3) and if the number of available function evaluations is about $100n$ or larger, several research directions emerge. The two main possibilities to integrate meta-models into the search with CMA-ES type algorithms are (i) the successive injection of the minimum of a learned meta-model at each time step into the learning of CMA-ES's covariance matrix and (ii) the use of a meta-model to predict the internal ranking of solutions. While for the latter, first results exist, the former idea is entirely unexplored for now. In both cases, a fundamental question is which type of meta-model (linear, quadratic, Gaussian Process, ...) is the best choice for a given number of function evaluations (as low as one or two function evaluations) and at which time the type of the meta-model shall be switched.

3.4. Setting novel standards in scientific experimentation and benchmarking

Numerical experimentation is needed as a complement to theory to test novel ideas, hypotheses, the stability of an algorithm, and/or to obtain quantitative estimates. Optimally, theory and experimentation go hand in hand, jointly guiding the understanding of the mechanisms underlying optimization algorithms. Though performing numerical experimentation on optimization algorithms is crucial and a common task, it is non-trivial and easy to fall in (common) pitfalls as stated by J. N. Hooker in his seminal paper [27].

In the RandOpt team we aim at raising the standards for both scientific experimentation and benchmarking.

[0] Often, this set forms a manifold of dimension one smaller than the number of objectives.

On the experimentation aspect, we are convinced that there is common ground over how scientific experimentation should be done across many (sub-)domains of optimization, in particular with respect to the visualization of results, testing extreme scenarios (parameter settings, initial conditions, etc.), how to conduct understandable and small experiments, how to account for invariance properties, performing scaling up experiments and so forth. We therefore want to formalize and generalize these ideas in order to make them known to the entire optimization community with the final aim that they become standards for experimental research.

Extensive numerical benchmarking, on the other hand, is a compulsory task for evaluating and comparing the performance of algorithms. It puts algorithms to a standardized test and allows to make recommendations which algorithms should be used preferably in practice. To ease this part of optimization research, we have been developing the Comparing Continuous Optimizers platform (COCO) since 2007 which allows to automatize the tedious task of benchmarking. It is a game changer in the sense that the freed time can now be spent on the scientific part of algorithm design (instead of implementing the experiments, visualization, statistical tests, etc.) and it opened novel perspectives in algorithm testing. COCO implements a thorough, well-documented methodology that is based on the above mentioned general principles for scientific experimentation.

Also due to the freely available data from 300+ algorithms benchmarked with the platform, COCO became a quasi-standard for single-objective, noiseless optimization benchmarking. It is therefore natural to extend the reach of COCO towards other subdomains (particularly constrained optimization, many-objective optimization) which can benefit greatly from an automated benchmarking methodology and standardized tests without (much) effort. This entails particularly the design of novel test suites and rethinking the methodology for measuring performance and more generally evaluating the algorithms. Particularly challenging is the design of scalable non-trivial testbeds for constrained optimization where one can still control where the solutions lies. Other optimization problem types, we are targeting are expensive problems (and the Bayesian optimization community in particular, see our AESOP project), optimization problems in machine learning (for example parameter tuning in reinforcement learning), and the collection of real-world problems from industry.

Another aspect of our future research on benchmarking is to investigate the large amounts of benchmarking data, we collected with COCO during the years. Extracting information about the influence of algorithms on the best performing portfolio, clustering algorithms of similar performance, or the automated detection of anomalies in terms of good/bad behavior of algorithms on a subset of the functions or dimensions are some of the ideas here.

Last, we want to expand the focus of COCO from automatized (large) benchmarking experiments towards everyday experimentation, for example by allowing the user to visually investigate algorithm internals on the fly or by simplifying the set up of algorithm parameter influence studies.

RAPSODI Project-Team

3. Research Program

3.1. Design and analysis of structure-preserving schemes

3.1.1. Numerical analysis of nonlinear numerical methods

Up to now, the numerical methods dedicated to degenerate parabolic problems that the mathematicians are able to analyze almost all rely on the use of mathematical transformations (like e.g. the Kirchhoff's transform). It forbids the extension of the analysis to complex realistic models. The methods used in the industrial codes for solving such complex problems rely on the use of what we call NNM, i.e., on methods that preserve all the nonlinearities of the problem without reducing them thanks to artificial mathematical transforms. Our aim is to take advantage of the recent breakthrough proposed by C. Cancès & C. Guichard [83], [4] to develop efficient new numerical methods with a full numerical analysis (stability, convergence, error estimates, robustness w.r.t. physical parameters,...).

3.1.2. Design and analysis of asymptotic-preserving schemes

There has been an extensive effort in the recent years to develop numerical methods for diffusion equations that are robust with respect to heterogeneities, anisotropy, and the mesh (see for instance [98] for an extensive discussion on such methods). On the other hand, the understanding of the role of nonlinear stability properties in the asymptotic behaviors of dissipative systems increased significantly in the last decades (see for instance [85], [110]).

Recently, C. Chainais-Hillairet and co-authors [79], [86] and [87] developed a strategy based on the control of the numerical counterpart of the physical entropy to develop and analyze AP numerical methods. In particular, these methods show great promises for capturing accurately the behavior of the solutions to dissipative problems when some physical parameter is small with respect to the discretization characteristic parameters, or in the long-time asymptotic. Since it requires the use of nonlinear test functions in the analysis, strong restrictions on the physics (isotropic problems) and on the mesh (Cartesian grids, Voronoï boxes...) are required in [79], [86] and [87]. The schemes proposed in [83] and [4] allow to handle nonlinear test functions in the analysis without restrictions on the mesh and on the anisotropy of the problem. Combining the nonlinear schemes *à la* [83] with the methodology of [79], [86], [87] would provide schemes that are robust both with respect to the meshes and to the parameters. Therefore, they would be also robust under adaptive mesh refinement.

3.1.3. Design and stability analysis of numerical methods for low-Mach models

We aim at extending the range of the `NS2DDV-M` software by introducing new physical models, like for instance the low-Mach model, which gives intermediate solutions between the compressible Navier–Stokes model and the incompressible Navier–Stokes one. This model was introduced in [109] as a limiting system which describes combustion processes at low Mach number in a confined region. Within this scope, we will propose a theoretical study for proving the existence of weak solutions for a particular class of models for which the dynamic viscosity of the fluid is a specific function of the density. We will propose also the extension of a combined Finite Volume-Finite Element method, initially developed for the simulation of incompressible and variable density flows, to this class of models.

3.2. Optimizing the computational efficiency

3.2.1. High-order nonlinear numerical methods

The numerical experiments carried out in [83] show that in case of very strong anisotropy, the convergence of the proposed NNM becomes too slow (less than first order). Indeed, the method appears to strongly overestimate the dissipation. In order to make the method more competitive, it is necessary to estimate the dissipation in a more accurate way. Preliminary numerical results show that second order accuracy in space can be achieved in this way. One also aims to obtain (at least) second order accuracy in time without jeopardizing the stability. For many problems, this can be done by using so-called two-step backward differentiation formulas (BDF2) [99].

Concerning the inhomogeneous fluid models, we aim to investigate new methods for the mass equation resolution. Indeed, we aim at increasing the accuracy while maintaining some positivity-like properties and the efficiency for a wide range of physical parameters. To this end, we will consider Residual Distribution schemes, that appear as an alternative to Finite Volume methods. Residual Distribution schemes enjoy very compact stencils. Therefore, their extension from 2D to 3D yield reasonable difficulties. These methods appeared twenty years ago, but recent extensions to unsteady problems [111], [106], with high-order accuracy [66], [65], or for parabolic problems [63], [64] make them very competitive. Relying on these breakthroughs, we aim at designing new Residual Distribution schemes for fluid mixture models with high-order accuracy while preserving the positivity of the solutions.

3.2.2. A posteriori error control

The question of the *a posteriori* error estimators will also have to be addressed in this optimization context. Since the pioneering papers of Babuska and Rheinboldt more than thirty years ago [70], *a posteriori* error estimators have been widely studied. We will take advantage of the huge corresponding bibliography database in order to optimize our numerical results.

For example, we would like to generalize the results we derived for the harmonic magnetodynamic case (e.g. [88] and [89]) to the temporal magnetodynamic one, for which space/time *a posteriori* error estimators have to be developed. A space/time refinement algorithm should consequently be proposed and tested on academic as well as industrial benchmarks.

We also want to develop *a posteriori* estimators for the variable density Navier–Stokes model or some of its variants. To do so, several difficulties have to be tackled: the problem is nonlinear, unsteady, and the numerical method [81], [82] we developed combines features from Finite Elements and Finite Volumes. Fortunately, we do not start from scratch. Some recent references are devoted to the unsteady Navier–Stokes model in the Finite Element context [77], [114]. In the Finite Volume context, recent references deal with unsteady convection-diffusion equations [113], [68], [96] and [84]. We want to adapt some of these results to the variable density Navier–Stokes system, and to be able to design an efficient space-time remeshing algorithm.

3.2.3. Efficient computation of pairwise interactions in large systems of particles

Many systems are modeled as a large number of punctual individuals (N) which interact pairwise which means $N(N-1)/2$ interactions. Such systems are ubiquitous, they are found in chemistry (Van der Waals interaction between atoms), in astrophysics (gravitational interactions between stars, galaxies or galaxy clusters), in biology (flocking behavior of birds, swarming of fishes) or in the description of crowd motions. Building on the special structure of convolution-type of the interactions, the team develops computation methods based on the Non Uniform Fast Fourier Transform [102]. This reduces the $O(N^2)$ naive computational cost of the interactions to $O(N \log N)$, allowing numerical simulations involving millions of individuals.

REALOPT Project-Team

3. Research Program

3.1. Introduction

integer programming, graph theory, decomposition approaches, polyhedral approaches, quadratic programming approaches, constraint programming.

Combinatorial optimization is the field of discrete optimization problems. In many applications, the most important decisions (control variables) are binary (on/off decisions) or integer (indivisible quantities). Extra variables can represent continuous adjustments or amounts. This results in models known as *mixed integer programs* (MIP), where the relationships between variables and input parameters are expressed as linear constraints and the goal is defined as a linear objective function. MIPs are notoriously difficult to solve: good quality estimations of the optimal value (bounds) are required to prune enumeration-based global-optimization algorithms whose complexity is exponential. In the standard approach to solving an MIP is so-called *branch-and-bound algorithm* : (i) one solves the linear programming (LP) relaxation using the simplex method; (ii) if the LP solution is not integer, one adds a disjunctive constraint on a factional component (rounding it up or down) that defines two sub-problems; (iii) one applies this procedure recursively, thus defining a binary enumeration tree that can be pruned by comparing the local LP bound to the best known integer solution. Commercial MIP solvers are essentially based on branch-and-bound (such IBM-CPLEX, FICO-Xpress-mp, or GUROBI). They have made tremendous progress over the last decade (with a speedup by a factor of 60). But extending their capabilities remains a continuous challenge; given the combinatorial explosion inherent to enumerative solution techniques, they remain quickly overwhelmed beyond a certain problem size or complexity.
Progress can be expected from the development of tighter formulations. Central to our field is the characterization of polyhedra defining or approximating the solution set and combinatorial algorithms to identify "efficiently" a minimum cost solution or separate an unfeasible point. With properly chosen formulations, exact optimization tools can be competitive with other methods (such as meta-heuristics) in constructing good approximate solutions within limited computational time, and of course has the important advantage of being able to provide a performance guarantee through the relaxation bounds. Decomposition techniques are implicitly leading to better problem formulation as well, while constraint propagation are tools from artificial intelligence to further improve formulation through intensive preprocessing. A new trend is robust optimization where recent progress have been made: the aim is to produce optimized solutions that remain of good quality even if the problem data has stochastic variations. In all cases, the study of specific models and challenging industrial applications is quite relevant because developments made into a specific context can become generic tools over time and see their way into commercial software.

Our project brings together researchers with expertise in mathematical programming (polyhedral approaches, decomposition and reformulation techniques in mixed integer programing, robust and stochastic programming, and dynamic programming), graph theory (characterization of graph properties, combinatorial algorithms) and constraint programming in the aim of producing better quality formulations and developing new methods to exploit these formulations. These new results are then applied to find high quality solutions for practical combinatorial problems such as routing, network design, planning, scheduling, cutting and packing problems, High Performance and Cloud Computing.

3.2. Polyhedral approaches for MIP

Adding valid inequalities to the polyhedral description of an MIP allows one to improve the resulting LP bound and hence to better prune the enumeration tree. In a cutting plane procedure, one attempt to identify valid inequalities that are violated by the LP solution of the current formulation and adds them to the formulation. This can be done at each node of the branch-and-bound tree giving rise to a so-called

branch-and-cut algorithm [65]. The goal is to reduce the resolution of an integer program to that of a linear program by deriving a linear description of the convex hull of the feasible solutions. Polyhedral theory tells us that if X is a mixed integer program: $X = P \cap \mathbb{Z}^n \times \mathbb{R}^p$ where $P = \{x \in \mathbb{R}^{n+p} : Ax \leq b\}$ with matrix $(A, b) \in \mathbb{Q}^{m \times (n+p+1)}$, then $conv(X)$ is a polyhedron that can be described in terms of linear constraints, i.e. it writes as $conv(X) = \{x \in \mathbb{R}^{n+p} : C\,x \leq d\}$ for some matrix $(C, d) \in \mathbb{Q}^{m' \times (n+p+1)}$ although the dimension m' is typically quite large. A fundamental result in this field is the equivalence of complexity between solving the combinatorial optimization problem $\min\{cx : x \in X\}$ and solving the *separation problem* over the associated polyhedron $conv(X)$: if $\widetilde{x} \notin conv(X)$, find a linear inequality $\pi\,x \geq \pi_0$ satisfied by all points in $conv(X)$ but violated by $\widetilde{x}$. Hence, for NP-hard problems, one can not hope to get a compact description of $conv(X)$ nor a polynomial time exact separation routine. Polyhedral studies focus on identifying some of the inequalities that are involved in the polyhedral description of $conv(X)$ and derive efficient *separation procedures* (cutting plane generation). Only a subset of the inequalities $C\,x \leq d$ can offer a good approximation, that combined with a branch-and-bound enumeration techniques permits to solve the problem. Using *cutting plane algorithm* at each node of the branch-and-bound tree, gives rise to the algorithm called *branch-and-cut*.

3.3. Decomposition-and-reformulation-approaches

An hierarchical approach to tackle complex combinatorial problems consists in considering separately different substructures (subproblems). If one is able to implement relatively efficient optimization on the substructures, this can be exploited to reformulate the global problem as a selection of specific subproblem solutions that together form a global solution. If the subproblems correspond to subset of constraints in the MIP formulation, this leads to Dantzig-Wolfe decomposition. If it corresponds to isolating a subset of decision variables, this leads to Bender's decomposition. Both lead to extended formulations of the problem with either a huge number of variables or constraints. Dantzig-Wolfe approach requires specific algorithmic approaches to generate subproblem solutions and associated global decision variables dynamically in the course of the optimization. This procedure is known as *column generation*, while its combination with branch-and-bound enumeration is called *branch-and-price*. Alternatively, in Bender's approach, when dealing with exponentially many constraints in the reformulation, the *cutting plane procedures* that we defined in the previous section are well-suited tools. When optimization on a substructure is (relatively) easy, there often exists a tight reformulation of this substructure typically in an extended variable space. This gives rise powerful reformulation of the global problem, although it might be impractical given its size (typically pseudo-polynomial). It can be possible to project (part of) the extended formulation in a smaller dimensional space if not the original variable space to bring polyhedral insight (cuts derived through polyhedral studies can often be recovered through such projections).

3.4. Integration of Artificial Intelligence Techniques in Integer Programming

When one deals with combinatorial problems with a large number of integer variables, or tightly constrained problems, mixed integer programming (MIP) alone may not be able to find solutions in a reasonable amount of time. In this case, techniques from artificial intelligence can be used to improve these methods. In particular, we use variable fixing techniques, primal heuristics and constraint programming.

Primal heuristics are useful to find feasible solutions in a small amount of time. We focus on heuristics that are either based on integer programming (rounding, diving, relaxation induced neighborhood search, feasibility pump), or that are used inside our exact methods (heuristics for separation or pricing subproblem, heuristic constraint propagation, ...). Such methods are likely to produce good quality solutions only if the integer programming formulation is of top quality, i.e., if its LP relaxation provides a good approximation of the IP solution.

In the same line, variable fixing techniques, that are essential in reducing the size of large scale problems, rely on good quality approximations: either tight formulations or tight relaxation solvers (as a dynamic program combined with state space relaxation). Then if the dual bound derives when the variable is fixed to one exceeds the incubent solution value, the variable can be fixed to zero and hence removed from the problem. The process can be apply sequentially by refining the degree of relaxation.

Constraint Programming (CP) focuses on iteratively reducing the variable domains (sets of feasible values) by applying logical and problem-specific operators. The latter propagates on selected variables the restrictions that are implied by the other variable domains through the relations between variables that are defined by the constraints of the problem. Combined with enumeration, it gives rise to exact optimization algorithms. A CP approach is particularly effective for tightly constrained problems, feasibility problems and min-max problems. Mixed Integer Programming (MIP), on the other hand, is known to be effective for loosely constrained problems and for problems with an objective function defined as the weighted sum of variables. Many problems belong to the intersection of these two classes. For such problems, it is reasonable to use algorithms that exploit complementary strengths of Constraint Programming and Mixed Integer Programming.

3.5. Robust Optimization

Decision makers are usually facing several sources of uncertainty, such as the variability in time or estimation errors. A simplistic way to handle these uncertainties is to overestimate the unknown parameters. However, this results in over-conservatism and a significant waste in resource consumption. A better approach is to account for the uncertainty directly into the decision aid model by considering mixed integer programs that involve uncertain parameters. Stochastic optimization account for the expected realization of random data and optimize an expected value representing the average situation. Robust optimization on the other hand entails protecting against the worst-case behavior of unknown data. There is an analogy to game theory where one considers an oblivious adversary choosing the realization that harms the solution the most. A full worst case protection against uncertainty is too conservative and induces very high over-cost. Instead, the realization of random data are bound to belong to a restricted feasibility set, the so-called uncertainty set. Stochastic and robust optimization rely on very large scale programs where probabilistic scenarios are enumerated. There is hope of a tractable solution for realistic size problems, provided one develops very efficient ad-hoc algorithms. The techniques for dynamically handling variables and constraints (column-and-row generation and Bender's projection tools) that are at the core of our team methodological work are specially well-suited to this context.

3.6. Approximation Algorithms

In some contexts, obtaining an exact solution to an optimization problem is not feasible: when instances are too large, or when decisions need to be taken rapidly. Since most of the combinatorial optimization problems are NP-hard, another direction to obtain good quality solutions in reasonable time is to focus on **approximation algorithms**. The definition of approximation algorithms is based on the notion of input set $\mathcal{I}$ and each $I \in \mathcal{I}$ defines a solution space $\mathcal{S}_I$. For a minimization problem $\min_{x \in \mathcal{S}_I} f(x)$, an algorithm $\mathcal{A}$ is an α-approximation algorithm if it provides a solution within α of the optimal solution for all instances in the input set:

$$\forall I \in \mathcal{I}, \quad f(\mathcal{A}(I)) \leq \alpha \min_{x \in \mathcal{S}_I} f(x) = f^*(I)$$

The objective is to search for polynomial algorithms, with approximation ratios as close to 1 as possible. Such algorithms are called *worst-case* approximation algorithms, because the performance guarantee is expressed over all possible inputs of the problem. The design of these algorithms have strong links with the enumeration techniques described above: since computing $f^*(I)$ is an NP-hard problem, it is often required to derive **strong a priori bounds** on the optimal solution value which can afterward be compared to estimations of the value of the solution produced. In many cases, it is also possible to build α-approximate solutions by a careful rounding of a solution obtained from the linear relaxation of an integer formulation of the problem. Members of the team have expertise in designing and evaluating approximation algorithms for resource allocation in computer systems, using a variety of techniques, such as dual approximation (where a guess of the optimal value f^* is provided, and $\mathcal{A}$ either provides a solution within αf^*, or guarantees that no solution of value f^* or less exists), or resource augmentation (where an approximation is obtained by relaxing some of the constraints of the problem).

3.7. Polyhedral Combinatorics and Graph Theory

Many fundamental combinatorial optimization problems can be modeled as the search for a specific structure in a graph. For example, ensuring connectivity in a network amounts to building a *tree* that spans all the nodes. Inquiring about its resistance to failure amounts to searching for a minimum cardinality *cut* that partitions the graph. Selecting disjoint pairs of objects is represented by a so-called *matching*. Disjunctive choices can be modeled by edges in a so-called *conflict graph* where one searches for *stable sets* – a set of nodes that are not incident to one another. Polyhedral combinatorics is the study of combinatorial algorithms involving polyhedral considerations. Not only it leads to efficient algorithms, but also, conversely, efficient algorithms often imply polyhedral characterizations and related min-max relations. Developments of polyhedral properties of a fundamental problem will typically provide us with more interesting inequalities well suited for a branch-and-cut algorithm to more general problems. Furthermore, one can use the fundamental problems as new building bricks to decompose the more general problem at hand. For problem that let themselves easily be formulated in a graph setting, the graph theory and in particular graph decomposition theorem might help.

SEQUEL Project-Team

3. Research Program

3.1. In Short

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical analysis and statistical learning, which provide the general concepts and tools to solve this problem.

To help the reader who is unfamiliar with these questions, we briefly present key ideas below.

3.2. Decision-making Under Uncertainty

The phrase "Decision under uncertainty" refers to the problem of taking decisions when we do not have a full knowledge neither of the situation, nor of the consequences of the decisions, as well as when the consequences of decision are non deterministic.

We introduce two specific sub-domains, namely the Markov decision processes which model sequential decision problems, and bandit problems.

3.2.1. Reinforcement Learning

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman's book [61].

A Markov Decision Process (MDP) is defined as the tuple $(\mathcal{X}, \mathcal{A}, P, r)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, P is the probabilistic transition kernel, and $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function. For the sake of simplicity, we assume in this introduction that the state and action spaces are finite. If the current state (at time t) is $x \in \mathcal{X}$ and the chosen action is $a \in \mathcal{A}$, then the Markov assumption means that the transition probability to a new state $x' \in \mathcal{X}$ (at time $t+1$) only depends on (x, a). We write $p(x'|x, a)$ the corresponding transition probability. During a transition $(x, a) \to x'$, a reward $r(x, a, x')$ is incurred.

In the MDP $(\mathcal{X}, \mathcal{A}, P, r)$, each initial state x_0 and action sequence $a_0, a_1, ...$ gives rise to a sequence of states $x_1, x_2, ...$, satisfying $\mathbb{P}(x_{t+1} = x'|x_t = x, a_t = a) = p(x'|x, a)$, and rewards [0]$r_1, r_2, ...$ defined by $r_t = r(x_t, a_t, x_{t+1})$.

The history of the process up to time t is defined to be $H_t = (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t)$. A policy π is a sequence of functions $\pi_0, \pi_1, ...$, where π_t maps the space of possible histories at time t to the space of probability distributions over the space of actions $\mathcal{A}$. To follow a policy means that, in each time step, we assume that the process history up to time t is $x_0, a_0, ..., x_t$ and the probability of selecting an action a is equal to $\pi_t(x_0, a_0, ..., x_t)(a)$. A policy is called stationary (or Markovian) if π_t depends only on the last visited state. In other words, a policy $\pi = (\pi_0, \pi_1, ...)$ is called stationary if $\pi_t(x_0, a_0, ..., x_t) = \pi_0(x_t)$ holds for all $t \geq 0$. A policy is called deterministic if the probability distribution prescribed by the policy for any history is concentrated on a single action. Otherwise it is called a stochastic policy.

[0]Note that for simplicity, we considered the case of a deterministic reward function, but in many applications, the reward r_t itself is a random variable.

We move from an MD process to an MD problem by formulating the goal of the agent, that is what the sought policy π has to optimize. It is very often formulated as maximizing (or minimizing), in expectation, some functional of the sequence of future rewards. For example, an usual functional is the infinite-time horizon sum of discounted rewards. For a given (stationary) policy π, we define the value function $V^\pi(x)$ of that policy π at a state $x \in \mathcal{X}$ as the expected sum of discounted future rewards given that we state from the initial state x and follow the policy π:

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, \pi\right], \tag{52}$$

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [60]) and average reward settings. Note also that, here, we consider the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [58], which introduces the optimal value function $V^*(x)$, defined as the optimal expected sum of rewards when the agent starts from a state x. We have $V^*(x) = \sup_\pi V^\pi(x)$. Now, let us give two definitions about policies:

- We say that a policy π is optimal, if it attains the optimal values $V^*(x)$ for any state $x \in \mathcal{X}$, *i.e.*, if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. Under mild conditions, deterministic stationary optimal policies exist [59]. Such an optimal policy is written π^*.
- We say that a (deterministic stationary) policy π is greedy with respect to (w.r.t.) some function V (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

$$\pi(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V(x')\right].$$

  where $\arg\max_{a \in \mathcal{A}} f(a)$ is the set of $a \in \mathcal{A}$ that maximizes $f(a)$. For any function V, such a greedy policy always exists because $\mathcal{A}$ is finite.

The goal of Reinforcement Learning (RL), as well as that of dynamic programming, is to design an optimal policy (or a good approximation of it).

The well-known Dynamic Programming equation (also called the Bellman equation) provides a relation between the optimal value function at a state x and the optimal value function at the successors states x' when choosing an optimal action: for all $x \in \mathcal{X}$,

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{53}$$

The benefit of introducing this concept of optimal value function relies on the property that, from the optimal value function V^*, it is easy to derive an optimal behavior by choosing the actions according to a policy greedy w.r.t. V^*. Indeed, we have the property that a policy greedy w.r.t. the optimal value function is an optimal policy:

$$\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{54}$$

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([64]):

- Bellman's dynamic programming approach, based on the introduction of the value function. It consists in learning a "good" approximation of the optimal value function, and then using it to derive a greedy policy w.r.t. this approximation. The hope (well justified in several cases) is that the performance V^{π} of the policy π greedy w.r.t. an approximation V of V^* will be close to optimality. This approximation issue of the optimal value function is one of the major challenges inherent to the reinforcement learning problem. **Approximate dynamic programming** addresses the problem of estimating performance bounds (*e.g.* the loss in performance $||V^* - V^{\pi}||$ resulting from using a policy π-greedy w.r.t. some approximation V- instead of an optimal policy) in terms of the approximation error $||V^* - V||$ of the optimal value function V^* by V. Approximation theory and Statistical Learning theory provide us with bounds in terms of the number of sample data used to represent the functions, and the capacity and approximation power of the considered function spaces.
- Pontryagin's maximum principle approach, based on sensitivity analysis of the performance measure w.r.t. some control parameters. This approach, also called **direct policy search** in the Reinforcement Learning community aims at directly finding a good feedback control law in a parameterized policy space without trying to approximate the value function. The method consists in estimating the so-called **policy gradient**, *i.e.* the sensitivity of the performance measure (the value function) w.r.t. some parameters of the current policy. The idea being that an optimal control problem is replaced by a parametric optimization problem in the space of parameterized policies. As such, deriving a policy gradient estimate would lead to performing a stochastic gradient method in order to search for a local optimal parametric policy.

Finally, many extensions of the Markov decision processes exist, among which the Partially Observable MDPs (POMDPs) is the case where the current state does not contain all the necessary information required to decide for sure of the best action.

3.2.2. Multi-arm Bandit Theory

Bandit problems illustrate the fundamental difficulty of decision making in the face of uncertainty: A decision maker must choose between what seems to be the best choice ("exploit"), or to test ("explore") some alternative, hoping to discover a choice that beats the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially. These bandit problems became popular with the seminal paper [62], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is specified by K real-valued distributions. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, *i.e.*, when the arm giving the highest expected reward is pulled all the time.

The name "bandit" comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which produces a random payoff as a result: When arm k is pulled, the random payoff is drawn from the distribution associated to k. Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, to play well, the gambler must carefully balance exploration and exploitation. Auer *et al.* [57] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most

at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

3.3. Statistical analysis of time series

Many of the problems of machine learning can be seen as extensions of classical problems of mathematical statistics to their (extremely) non-parametric and model-free cases. Other machine learning problems are founded on such statistical problems. Statistical problems of sequential learning are mainly those that are concerned with the analysis of time series. These problems are as follows.

3.3.1. Prediction of Sequences of Structured and Unstructured Data

Given a series of observations $x_1, ..., x_n$ it is required to give forecasts concerning the distribution of the future observations $x_{n+1}, x_{n+2}, \cdots$; in the simplest case, that of the next outcome x_{n+1}. Then x_{n+1} is revealed and the process continues. Different goals can be formulated in this setting. One can either make some assumptions on the probability measure that generates the sequence $x_1, \cdots, x_n, \cdots$, such as that the outcomes are independent and identically distributed (i.i.d.), or that the sequence is a Markov chain, that it is a stationary process, etc. More generally, one can assume that the data is generated by a probability measure that belongs to a certain set $\mathcal{C}$. In these cases the goal is to have the discrepancy between the predicted and the "true" probabilities to go to zero, if possible, with guarantees on the speed of convergence.

Alternatively, rather than making some assumptions on the data, one can change the goal: the predicted probabilities should be asymptotically as good as those given by the best reference predictor from a certain pre-defined set.

Another dimension of complexity in this problem concerns the nature of observations x_i. In the simplest case, they come from a finite space, but already basic applications often require real-valued observations. Moreover, function or even graph-valued observations often arise in practice, in particular in applications concerning Web data. In these settings estimating even simple characteristics of probability distributions of the future outcomes becomes non-trivial, and new learning algorithms for solving these problems are in order.

3.3.2. Hypothesis testing

Given a series of observations of $x_1, \cdots, x_n, \cdots$ generated by some unknown probability measure μ, the problem is to test a certain given hypothesis H_0 about μ, versus a given alternative hypothesis H_1. There are many different examples of this problem. Perhaps the simplest one is testing a simple hypothesis "μ is Bernoulli i.i.d. measure with probability of 0 equals 1/2" versus "μ is Bernoulli i.i.d. with the parameter different from 1/2". More interesting cases include the problems of model verification: for example, testing that μ is a Markov chain, versus that it is a stationary ergodic process but not a Markov chain. In the case when we have not one but several series of observations, we may wish to test the hypothesis that they are independent, or that they are generated by the same distribution. Applications of these problems to a more general class of machine learning tasks include the problem of feature selection, the problem of testing that a certain behavior (such as pulling a certain arm of a bandit, or using a certain policy) is better (in terms of achieving some goal, or collecting some rewards) than another behavior, or than a class of other behaviors.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis H_0 and H_1 about the unknown measure that generates the data, find out whether it is possible to test H_0 against H_1 (with confidence), and if so, how can one do it.

3.3.3. Change Point Analysis

A stochastic process is generating the data. At some point, the process distribution changes. In the "offline" situation, the statistician observes the resulting sequence of outcomes and has to estimate the point or the points at which the change(s) occurred. In online setting, the goal is to detect the change as quickly as possible.

These are the classical problems in mathematical statistics, and probably among the last remaining statistical problems not adequately addressed by machine learning methods. The reason for the latter is perhaps in that the problem is rather challenging. Thus, most methods available so far are parametric methods concerning piece-wise constant distributions, and the change in distribution is associated with the change in the mean. However, many applications, including DNA analysis, the analysis of (user) behavior data, etc., fail to comply with this kind of assumptions. Thus, our goal here is to provide completely non-parametric methods allowing for any kind of changes in the time-series distribution.

3.3.4. Clustering Time Series, Online and Offline

The problem of clustering, while being a classical problem of mathematical statistics, belongs to the realm of unsupervised learning. For time series, this problem can be formulated as follows: given several samples $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by k different distributions.

The online version of the problem allows for the number of observed time series to grow with time, in general, in an arbitrary manner.

3.3.5. Online Semi-Supervised Learning

Semi-supervised learning (SSL) is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is extremely useful for solving real-world problems, where data is often abundant but the resources to label them are limited.

Furthermore, *online* SSL is suitable for adaptive machine learning systems. In the classification case, learning is viewed as a repeated game against a potentially adversarial nature. At each step t of this game, we observe an example $\mathbf{x_t}$, and then predict its label $\widehat{y}_t$.

The challenge of the game is that we only exceptionally observe the true label y_t. In the extreme case, which we also study, only a handful of labeled examples are provided in advance and set the initial bias of the system while unlabeled examples are gathered online and update the bias continuously. Thus, if we want to adapt to changes in the environment, we have to rely on indirect forms of feedback, such as the structure of data.

3.3.6. Online Kernel and Graph-Based Methods

Large-scale kernel ridge regression is limited by the need to store a large kernel matrix. Similarly, large-scale graph-based learning is limited by storing the graph Laplacian. Furthermore, if the data come online, at some point no finite storage is sufficient and per step operations become slow.

Our challenge is to design sparsification methods that give guaranteed approximate solutions with a reduced storage requirements.

SIERRA Project-Team

3. Research Program

3.1. Supervised Learning

This part of our research focuses on methods where, given a set of examples of input/output pairs, the goal is to predict the output for a new input, with research on kernel methods, calibration methods, and multi-task learning.

3.2. Unsupervised Learning

We focus here on methods where no output is given and the goal is to find structure of certain known types (e.g., discrete or low-dimensional) in the data, with a focus on matrix factorization, statistical tests, dimension reduction, and semi-supervised learning.

3.3. Parsimony

The concept of parsimony is central to many areas of science. In the context of statistical machine learning, this takes the form of variable or feature selection. The team focuses primarily on structured sparsity, with theoretical and algorithmic contributions.

3.4. Optimization

Optimization in all its forms is central to machine learning, as many of its theoretical frameworks are based at least in part on empirical risk minimization. The team focuses primarily on convex and bandit optimization, with a particular focus on large-scale optimization.

SIMSMART Project-Team

3. Research Program

3.1. Research Program

Introduction. Computer simulation of physical systems is becoming increasingly reliant on highly complex models, as the constant surge of computational power is nurturing scientists into simulating the most detailed features of reality – from complex molecular systems to climate/weather forecast.

Yet, when modeling physical reality, bottom-up approaches are stumbling over intrinsic difficulties. First, the timescale separation between the fastest simulated microscopic features, and the macroscopic effective slow behavior becomes huge, implying that the fully detailed and direct long time simulation of many interesting systems (*e.g.* large molecular systems) are out of reasonable computational reach. Second, the chaotic dynamical behaviors of the systems at stake, coupled with such multi-scale structures, exacerbate the intricate uncertainty of outcomes, which become highly dependent on intrinsic chaos, uncontrolled modeling, as well as numerical discretization. Finally, the massive increase of observational data addresses new challenges to classical data assimilation, such as dealing with high dimensional observations and/or extremely long time series of observations.

SIMSMART Identity. Within this highly challenging applicative context, SIMSMART positions itself as a computational probability and statistics research team, with a mathematical perspective. Our approach is based on the use of *stochastic modeling* of complex physical systems, and on the use of *Monte Carlo simulation* methods, with a strong emphasis on dynamical models. The two main numerical tasks of interest to SIMSMART are the following: (i) simulating with pseudo-random number generators - a.k.a. *sampling* - dynamical models of random physical systems, (ii) sampling such random physical dynamical models given some real observations - a.k.a. *Bayesian data assimilation*. SIMSMART aims at providing an appropriate mathematical level of abstraction and generalization to a wide variety of Monte Carlo simulation algorithms in order to propose non-superficial answers to both *methodological and mathematical* challenges. The issues to be resolved include computational complexity reduction, statistical variance reduction, and uncertainty quantification.

SIMSMART's Objectives. The main objective of SIMSMART is to disrupt this now classical field of particle Monte Carlo simulation by creating deeper mathematical frameworks adapted to the challenging world of complex (*e.g.* high dimensional and/or multi-scale), and massively observed systems, as described in the beginning of this introduction.

To be more specific, we will classify SIMSMART objectives using the following four intertwined topics:

1. Objective 1: Rare events and random simulation.
2. Objective 2: High dimensional and advanced particle filtering.
3. Objective 3: Non-parametric approaches.
4. Objective 4: Model reduction and sparsity.

Rare events Objective 1 are ubiquitous in random simulation, either to accelerate the occurrence of physically relevant random slow phenomenons, or to estimate the effect of uncertain variables. Objective 1 will be mainly concerned with particle methods where *splitting* is used to enforce the occurrence of rare events.

The problem of high dimensional observations, the main topic in Objective 2, is a known bottleneck in filtering, especially in non-linear particle filtering, where linear data assimilation methods remain the state-of-the-art approaches.

The increasing size of recorded observational data and the increasing complexity of models also suggest to devote more effort into non-parametric data assimilation methods, the main issue of Objective 3.

In some contexts, for instance when one wants to compare solutions of a complex (*e.g.* high dimensional) dynamical systems depending on uncertain parameters, the construction of relevant reduced-order models becomes a key topic. This is the content of Objective 4.

With respect to volume of research activity, Objective 1, Objective 4 and the sum (Objective 2+Objective 3) are comparable.

Some new challenges in the simulation and data assimilation of random physical dynamical systems have become prominent in the last decade. A first issue (i) consists in the intertwined problems of simulating on large, macroscopic random times, and simulating *rare events*. The link between both aspects stems from the fact that many effective, large times dynamics can be approximated by sequences of rare events. A second, obvious, issue (ii) consists in managing *very abundant observational data*. A third issue (iii) consists in quantifying *uncertainty/sensitivity/variance* of outcomes with respect to models or noise. A fourth issue (iv) consists in managing *high dimensionality*, either when dealing with complex prior physical models, or with very large data sets. The related increase of complexity also requires, as a fifth issue (v), the construction of *reduced models* to speed-up comparative simulations. In a context of very abundant data, this may be replaced by a sixth issue (vi) where complexity constraints on modeling is replaced by the use of *non-parametric statistical inference*.

Hindsight suggests that all the latter challenges are related. Indeed, the contemporary digital condition, made of a massive increase in computational power and in available data, is resulting in a demand for more complex and uncertain models, for more extreme regimes, and for using inductive approaches relying on abundant data.

For simplicity, we have classified SIMSMART research into the following already mentioned four main objectives.

1. Objective 1: Rare events and random simulation, which mainly encompass item (i).
2. Objective 2: High dimension and advanced particle filtering, which encompass item (iv).
3. Objective 3: Non-parametric inference, which mainly encompass item (ii) and (vi).
4. Objective 4: Model reduction, which mainly encompasses item (vi).

Uncertainty quantification (item (iii)) in fact underlies each aspect since we are mainly interested in Monte Carlo approaches, so that uncertainty can be *modeled by an initial random variable and be incorporated in the state space of the physical model*.

SPHINX Project-Team

3. Research Program

3.1. Control and stabilization of heterogeneous systems

Fluid-Structure Interaction Systems (FSIS) are present in many physical problems and applications. Their study involves solving several challenging mathematical problems:

- **Nonlinearity:** One has to deal with a system of nonlinear PDE such as the Navier-Stokes or the Euler systems;
- **Coupling:** The corresponding equations couple two systems of different types and the methods associated with each system need to be suitably combined to solve successfully the full problem;
- **Coordinates:** The equations for the structure are classically written with Lagrangian coordinates whereas the equations for the fluid are written with Eulerian coordinates;
- **Free boundary:** The fluid domain is moving and its motion depends on the motion of the structure. The fluid domain is thus an unknown of the problem and one has to solve a free boundary problem.

In order to control such FSIS systems, one has first to analyze the corresponding system of PDE. The oldest works on FSIS go back to the pioneering contributions of Thomson, Tait and Kirchhoff in the 19th century and Lamb in the 20th century, who considered simplified models (potential fluid or Stokes system). The first mathematical studies in the case of a viscous incompressible fluid modeled by the Navier-Stokes system and a rigid body whose dynamics is modeled by Newton's laws appeared much later [119], [114], [94], and almost all mathematical results on such FSIS have been obtained in the last twenty years.

The most studied FSIS is the problem modeling a **rigid body moving in a viscous incompressible fluid** ([77], [73], [112], [83], [88], [116], [118], [102], [86]). Many other FSIS have been studied as well. Let us mention [104], [91], [87], [76], [64], [82], [65], [84] for different fluids. The case of **deformable structures** has also been considered, either for a fluid inside a moving structure (e.g. blood motion in arteries) or for a moving deformable structure immersed in a fluid (e.g. fish locomotion). The obtained coupled FSIS is a complex system and its study raises several difficulties. The main one comes from the fact that we gather two systems of different nature. Some studies have been performed for approximations of this system: [69], [64], [97], [78], [67]). Without approximations, the only known results [74], [75] were obtained with very strong assumptions on the regularity of the initial data. Such assumptions are not satisfactory but seem inherent to this coupling between two systems of different natures. In order to study self-propelled motions of structures in a fluid, like fish locomotion, one can assume that the **deformation of the structure is prescribed and known**, whereas its displacement remains unknown ([110]). This permits to start the mathematical study of a challenging problem: understanding the locomotion mechanism of aquatic animals. This is related to control or stabilization problems for FSIS. Some first results in this direction were obtained in [92], [66], [106].

3.2. Inverse problems for heterogeneous systems

The area of inverse problems covers a large class of theoretical and practical issues which are important in many applications (see for instance the books of Isakov [93] or Kaltenbacher, Neubauer, and Scherzer [95]). Roughly speaking, an inverse problem is a problem where one attempts to recover an unknown property of a given system from its response to an external probing signal. For systems described by evolution PDE, one can be interested in the reconstruction from partial measurements of the state (initial, final or current), the inputs (a source term, for instance) or the parameters of the model (a physical coefficient for example). For stationary or periodic problems (i.e. problems where the time dependence is given), one can be interested in determining from boundary data a local heterogeneity (shape of an obstacle, value of a physical coefficient describing the medium, etc.). Such inverse problems are known to be generally ill-posed and their study leads to investigate the following questions:

- *Uniqueness.* The question here is to know whether the measurements uniquely determine the unknown quantity to be recovered. This theoretical issue is a preliminary step in the study of any inverse problem and can be a hard task.
- *Stability.* When uniqueness is ensured, the question of stability, which is closely related to sensitivity, deserves special attention. Stability estimates provide an upper bound for the parameter error given some uncertainty on data. This issue is closely related to the so-called observability inequality in systems theory.
- *Reconstruction.* Inverse problems being usually ill-posed, one needs to develop specific reconstruction algorithms which are robust to noise, disturbances and discretization. A wide class of methods is based on optimization techniques.

We can split our research in inverse problems into two classes which both appear in FSIS and CWS:

1. **Identification for evolution PDE.**

   Driven by applications, the identification problem for systems of infinite dimension described by evolution PDE has seen in the last three decades a fast and significant growth. The unknown to be recovered can be the (initial/final) state (e.g. state estimation problems [59], [85], [89], [115] for the design of feedback controllers), an input (for instance source inverse problems [56], [68], [79]) or a parameter of the system. These problems are generally ill-posed and many regularization approaches have been developed. Among the different methods used for identification, let us mention optimization techniques ([72]), specific one-dimensional techniques (like in [60]) or observer-based methods as in [100].

   In the last few years, we have developed observers to solve initial data inverse problems for a class of linear systems of infinite dimension. Let us recall that observers, or Luenberger observers [99], have been introduced in automatic control theory to estimate the state of a dynamical system of finite dimension from the knowledge of an output (for more references, see for instance [103] or [117]). Using observers, we have proposed in [105], [90] an iterative algorithm to reconstruct initial data from partial measurements for some evolution equations.We are deepening our activities in this direction by considering more general operators or more general sources and the reconstruction of coefficients for the wave equation. In connection with this problem, we study the stability in the determination of these coefficients. To achieve this, we use geometrical optics, which is a classical albeit powerful tool to obtain quantitative stability estimates on some inverse problems with a geometrical background, see for instance [62], [61].

2. **Geometric inverse problems.**

   We investigate some geometric inverse problems that appear naturally in many applications, like medical imaging and non destructive testing. A typical problem we have in mind is the following: given a domain Ω containing an (unknown) local heterogeneity ω, we consider the boundary value problem of the form

$$\begin{cases} Lu = 0, & (\Omega \setminus \omega) \\ u = f, & (\partial\Omega) \\ Bu = 0, & (\partial\omega) \end{cases}$$

   where L is a given partial differential operator describing the physical phenomenon under consideration (typically a second order differential operator), B the (possibly unknown) operator describing the boundary condition on the boundary of the heterogeneity and f the exterior source used to probe the medium. The question is then to recover the shape of ω and/or the boundary operator B from some measurement Mu on the outer boundary $\partial\Omega$. This setting includes in particular inverse scattering problems in acoustics and electromagnetics (in this case Ω is the whole space and the data are far

field measurements) and the inverse problem of detecting solids moving in a fluid. It also includes, with slight modifications, more general situations of incomplete data (i.e. measurements on part of the outer boundary) or penetrable inhomogeneities. Our approach to tackle this type of problems is based on the derivation of a series expansion of the input-to-output map of the problem (typically the Dirichlet-to-Neumann map of the problem for the Calderón problem) in terms of the size of the obstacle.

3.3. Numerical analysis and simulation of heterogeneous systems

Within the team, we have developed in the last few years numerical codes for the simulation of FSIS and CWS. We plan to continue our efforts in this direction.

- In the case of FSIS, our main objective is to provide computational tools for the scientific community, essentially to solve academic problems.
- In the case of CWS, our main objective is to build tools general enough to handle industrial problems. Our strong collaboration with Christophe Geuzaine's team in Liège (Belgium) makes this objective credible, through the combination of DDM (Domain Decomposition Methods) and parallel computing.

Below, we explain in detail the corresponding scientific program.

- **Simulation of FSIS:** In order to simulate fluid-structure systems, one has to deal with the fact that the fluid domain is moving and that the two systems for the fluid and for the structure are strongly coupled. To overcome this free boundary problem, three main families of methods are usually applied to numerically compute in an efficient way the solutions of the fluid-structure interaction systems. The first method consists in suitably displacing the mesh of the fluid domain in order to follow the displacement and the deformation of the structure. A classical method based on this idea is the A.L.E. (Arbitrary Lagrangian Eulerian) method: with such a procedure, it is possible to keep a good precision at the interface between the fluid and the structure. However, such methods are difficult to apply for large displacements (typically the motion of rigid bodies). The second family of methods consists in using a *fixed mesh* for both the fluid and the structure and to simultaneously compute the velocity field of the fluid with the displacement velocity of the structure. The presence of the structure is taken into account through the numerical scheme. Finally, the third class of methods consists in transforming the set of PDEs governing the flow into a system of integral equations set on the boundary of the immersed structure. The members of SPHINX have already worked on these three families of numerical methods for FSIS systems with rigid bodies (see e.g. [109], [96], [111], [107], [108], [101]).
- **Simulation of CWS:** Solving acoustic or electromagnetic scattering problems can become a tremendously hard task in some specific situations. In the high frequency regime (i.e. for small wavelength), acoustic (Helmholtz's equation) or electromagnetic (Maxwell's equations) scattering problems are known to be difficult to solve while being crucial for industrial applications (e.g. in aeronautics and aerospace engineering). Our particularity is to develop new numerical methods based on the hybridization of standard numerical techniques (like algebraic preconditioners, etc.) with approaches borrowed from asymptotic microlocal analysis. Most particularly, we contribute to building hybrid algebraic/analytical preconditioners and quasi-optimal Domain Decomposition Methods (DDM) [63], [80], [81] for highly indefinite linear systems. Corresponding three-dimensional solvers (like for example GetDDM) will be developed and tested on realistic configurations (e.g. submarines, complete or parts of an aircraft, etc.) provided by industrial partners (Thales, Airbus). Another situation where scattering problems can be hard to solve is the one of dense multiple (acoustic, electromagnetic or elastic) scattering media. Computing waves in such media requires us to take into account not only the interactions between the incident wave and the scatterers, but also the effects of the interactions between the scatterers themselves. When the number of scatterers is very large (and possibly at high frequency [58], [57]), specific deterministic or stochastic numerical methods and algorithms are needed. We introduce new optimized numerical methods for solving such complex

configurations. Many applications are related to this problem *e.g.* for osteoporosis diagnosis where quantitative ultrasound is a recent and promising technique to detect a risk of fracture. Therefore, numerical simulation of wave propagation in multiple scattering elastic media in the high frequency regime is a very useful tool for this purpose.

TAU Project-Team

3. Research Program

3.1. Toward Good AI

As discussed by [141], the topic of ethical AI was non-existent until 2010, was laughed at in 2016, and became a hot topic in 2017 as the AI disruptivity with respect to the fabric of life (travel, education, entertainment, social networks, politics, to name a few) became unavoidable [138], together with its expected impacts on the nature and amount of jobs. As of now, it seems that the risk of a new AI Winter might arise from legal[0] and societal[0] issues. While privacy is now recognized as a civil right in Europe, it is feared that the GAFAM, BATX and others can already capture a sufficient fraction of human preferences and their dynamics to achieve their commercial and other goals, and build a Brave New Big Brother (BNBB, a system that is openly beneficial to many, covertly nudging, and possibly dictatorial).

The ambition of TAU is to mitigate the BNBB risk along several intricated dimensions, and build i) causal and explainable models; ii) fair data and models; iii) provably robust models.

3.1.1. Causal modeling and biases

Participants: Isabelle Guyon, Michèle Sebag, Philippe Caillou, Paola Tubaro
PhD: Diviyan Kalainathan
Collaboration: Olivier Goudet (Université d'Angers), David Lopez-Paz (Facebook)

The extraction of causal models, a long goal of AI [139], [117], [140], became a strategic issue as the usage of learned models gradually shifted from *prediction* to *prescription* in the last years. This evolution, following Auguste Comte's vision of science (*Savoir pour prévoir, afin de pouvoir*) indeed reflects the exuberant optimism about AI: Knowledge enables Prediction; Prediction enables Control. However, although predictive models can be based on correlations, prescriptions can only be based on causal models[0].

Among the research applications concerned with causal modeling, predictive modeling or collaborative filtering at TAU are all projects described in section 4.1 (see also Section 3.4), studying the relationships between: i) the educational background of persons and the job openings (FUI project JobAgile and DataIA project Vadore); ii) the quality of life at work and the economic performance indicators of the enterprises (ISN Lidex project Amiqap) [119] ; iii) the nutritional items bought by households (at the level of granularity of the barcode) and their health status, as approximated from their body-mass-index (IRS UPSaclay Nutriperso); iv) the actual offer of restaurants and their scores on online rating systems. In these projects, a wealth of data is available (though hardly sufficient for applications ii), iii and iv))) and there is little doubt that these data reflect the imbalances and biases of the world as is, ranging from gender to racial to economical prejudices. Preventing the learned models from perpetuating such biases is essential to deliver an AI endowed with common decency.

In some cases, the bias is known; for instance, the cohorts in the Nutriperso study are more well-off than the average French population, and the Kantar database includes explicit weights to address this bias through importance sampling. In other cases, the bias is only guessed; for instance, the companies for which Secafi data are available hardly correspond to a uniform sample as these data have been gathered upon the request of the company trade union.

[0]For instance, the (fictitious) plea challenge proposed to law students in Oct. 2018 considered a chain reaction pileup occurred among autonomous and humanly operated vehicles on a highway.
[0]For instance related to information bubbles and nudge [100], [155].
[0]One can predict that it rains based on the presence of umbrellas in the street; but one cannot induce rainfall by going out with an umbrella. Likewise, the presence of books/tablets at home and the good scores of children at school are correlated; but offering books/tablets to all children might fail to improve their scores *per se*, if both good scores and books are explained by a so-called confounder variable, like the presence of adults versed in books/tablets at home.

3.1.2. Robustness of Learned Models

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag
PhDs: Julien Girard, Marc Nabhan, Nizham Makhoud
Collaboration: Zakarian Chihani (CEA); Hiba Hage, and Yves Tourbier (Renault); Jérôme Kodjabachian (Thalès THERESIS)

Due to their outstanding performances, deep neural networks and more generally machine learning-based decision making systems, referred to as MLs in the following, have been raising hopes in the recent years to achieve breakthroughs in critical systems, ranging from autonomous vehicles to defense. The main pitfall for such applications lies in the lack of guarantees for MLs robustness.

Specifically, MLs are used when the mainstream software design process does not apply, that is, when no formal specification of the target software behavior is available and/or when the system is embedded in an open unpredictable world. The extensive body of knowledge developed to deliver guarantees about mainstream software – ranging from formal verification, model checking and abstract interpretation to testing, simulation and monitoring – thus does not directly apply either. Another weakness of MLs regards their dependency to the amount and quality of the training data, as their performances are sensitive to slight perturbations of the data distribution. Such perturbations can occur naturally due to domain or concept drift (e.g. due to a change in light intensity or a scratch on a camera lens); they can also result from intentional malicious attacks, a.k.a adversarial examples [156].

These downsides, currently preventing the dissemination of MLs in safety-critical systems (SCS), call for a considerable amount of research, in order to understand when and to which extent an MLs can be certified to provide the desired level of guarantees.

Julien Girard's PhD (CEA scholarship), started in Oct. 2018, co-supervised by Guillaume Charpiat and Zakaria Chihani (CEA), is devoted to the extension of abstract interpretation to deep neural nets, and the formal characterization of the transition kernel from input to output space achieved by a DNN (robustness by design, coupled with formally assessing the coverage of the training set). This approach is tightly related to the inspection and opening of black-box models, aimed to characterize the patterns in the input instances responsible for a decision – another step toward explainability.

On the other hand, experimental validation of MLs, akin statistical testing, also faces three limitations: i) real-world examples are notoriously insufficient to ensure a good coverage in general; ii) for this reason, simulated examples are extensively used; but their use raises the *reality gap* issue [128] of the distance between real and simulated worlds; iii) independently, the real-world is naturally subject to domain shift (e.g. due to the technical improvement and/or aging of sensors). Our collaborations with Renault tackle such issues in the context of the autonomous vehicle (see Section 7.1.3).

3.2. Hybridizing numerical modeling and learning systems

Participants: Alessandro Bucci, Guillaume Charpiat, Cécile Germain, Isabelle Guyon, Marc Schoenauer, Michèle Sebag
PhD: Théophile Sanchez, Loris Felardos, Wenzhuo Liu
In sciences and engineering, human knowledge is commonly expressed in closed form, through equations or mechanistic models characterizing how a natural or social phenomenon, or a physical device, will behave/evolve depending on its environment and external stimuli, under some assumptions and up to some approximations. The field of numerical engineering, and the simulators based on such mechanistic models, are at the core of most approaches to understand and analyze the world, from solid mechanics to computational fluid dynamics, from chemistry to molecular biology, from astronomy to population dynamics, from epidemiology and information propagation in social networks to economy and finance.

Most generally, numerical engineering supports the simulation, and when appropriate the optimization and control[0] of the phenomenons under study, although several sources of discrepancy might adversely affect the results, ranging from the underlying assumptions and simplifying hypotheses in the models, to systematic experiment errors to statistical measurement errors (not to mention numerical issues). This knowledge and know-how are materialized in millions of lines of code, capitalizing the expertise of academic and industrial labs. These softwares have been steadily extended over decades, modeling new and more fine-grained effects through layered extensions, making them increasingly harder to maintain, extend and master. Another difficulty is that complex systems most often resort to hybrid (pluridisciplinary) models, as they involve many components interacting along several time and space scales, hampering their numerical simulation.

At the other extreme, machine learning offers the opportunity to model phenomenons from scratch, using any available data gathered through experiments or simulations. Recent successes of machine learning in computer vision, natural language processing and games to name a few, have demonstrated the power of such agnostic approaches and their efficiency in terms of prediction [123], inverse problem solving [170], and sequential decision making [162], [81], despite their lack of any "semantic" understanding of the universe. Even before these successes, Anderson's claim was that *the data deluge [might make] the scientific method obsolete* [70], as if a reasonable option might be to throw away the existing equational or software bodies of knowledge, and let Machine Learning rediscover all models from scratch. Such a claim is hampered among others by the fact that not all domains offer a wealth of data, as any academic involved in an industrial collaboration around data has discovered.

Another approach will be considered in TAU, investigating how existing mechanistic models and related simulators can be partnered with ML algorithms: i) to achieve the same goals with the same methods with a gain of accuracy or time; ii) to achieve new goals; iii) to achieve the same goals with new methods.

Toward more robust numerical engineering: In domains where satisfying mechanistic models and simulators are available, ML can contribute to improve their accuracy or usability. A first direction is to refine or extend the models and simulators to better fit the empirical evidence. The goal is to finely account for the different biases and uncertainties attached to the available knowledge and data, distinguishing the different types of *known unknowns*. Such *known unknowns* include the model hyper-parameters (coefficients), the systematic errors due to e.g., experiment imperfections, and the statistical errors due to e.g., measurement errors. A second approach is based on learning a surrogate model for the phenomenon under study that incorporate domain knowledge from the mechanistic model (or its simulation). See Section 7.5 for case studies.

A related direction, typically when considering black-box simulators, aims to learn a model of the error, or equivalently, a post-processor of the software. The discrepancy between simulated and empirical results, referred to as *reality gap* [128], can be tackled in terms of domain adaptation [74], [99]. Specifically, the source domain here corresponds to the simulated phenomenon, offering a wealth of inexpensive data, and the target domain corresponds to the actual phenomenon, with rare and expensive data; the goal is to devise accurate target models using the source data and models.

Extending numerical engineering: ML, using both experimental and numerical data, can also be used to tackle new goals, that are beyond the current state-of-the-art of standard approaches. Inverse problems are such goals, identifying the parameters or the initial conditions of phenomenons for which the model is not differentiable, or amenable to the adjoint state method.

A slightly different kind of inverse problem is that of recovering the ground truth when only noisy data is available. This problem can be formulated as a search for the simplest model explaining the data. The question then becomes to formulate and efficiently exploit such a simplicity criterion.

Another goal can be to model the distribution of given quantiles for some system: The challenge is to exploit available data to train a generative model, aimed at sampling the target quantiles.

[0]Note that the causal nature of mechanistic models is established from prior knowledge and experimentations.

Examples tackled in TAU are detailed in Section 7.5 . Note that the "Cracking the Glass Problem", described in Section 7.2.3 is yet another instance of a similar problem.

Data-driven numerical engineering : Finally, ML can also be used to sidestep numerical engineering limitations in terms of scalability, or to build a simulator emulating the resolution of the (unknown) mechanistic model from data, or to revisit the formal background.

When the mechanistic model is known and sufficiently accurate, it can be used to train a deep network on an arbitrary set of (space,time) samples, resulting in a meshless numerical approximation of the model [151], supporting by construction *differentiable programming* [125].

When no mechanistic model is sufficiently efficient, the model must be identified from the data only. Genetic programming has been used to identify systems of ODEs [149], through the identification of invariant quantities from data, as well as for the direct identification of control commands of nonlinear complex systems, including some chaotic systems [88]. Another recent approach uses two deep neural networks, one for the state of the system, the other for the equation itself [142]. The critical issues for both approaches include the scalability, and the explainability of the resulting models. Such line of research will benefit from TAU unique mixed expertise in Genetic Programming and Deep Learning.

Finally, in the realm of signal processing (SP), the question is whether and how deep networks can be used to revisit mainstream feature extraction based on Fourier decomposition, wavelet and scattering transforms [76]. E. Bartenlian's PhD (started Oct. 2018), co-supervised by M. Sebag and F. Pascal (Centrale-Supélec), focusing on musical audio-to-score translation [150], inspects the effects of supervised training, taking advantage from the fact that convolution masks can be initialized and analyzed in terms of frequency.

3.3. Learning to learn

According to Ali Rahimi's test of times award speech at NIPS 17, the current ML algorithms *have become a form of alchemy*. Competitive testing and empirical breakthroughs gradually become mandatory for a contribution to be acknowledged; an increasing part of the community adopts trials and errors as main scientific methodology, and theory is lagging behind practice. This style of progress is typical of technological and engineering revolutions for some; others ask for consolidated and well-understood theoretical advances, saving the time wasted in trying to build upon hardly reproducible results.

Basically, while practical achievements have often passed the expectations, there exist caveats along three dimensions. Firstly, excellent performances do not imply that the model has captured what was to learn, as shown by the phenomenon of adversarial examples. Following Ian Goodfellow, some well-performing models might be compared to *Clever Hans*, the horse that was able to solve mathematical exercizes using non verbal cues from its teacher [116]; it is the purpose of Pillar I. to alleviate the *Clever Hans* trap (section 3.1).
Secondly, some major advances, e.g. related to the celebrated adversarial learning [105], [99], establish proofs of concept more than a sound methodology, where the reproducibility is limited due to i) the computational power required for training (often beyond reach of academic labs); ii) the numerical instabilities (witnessed as random seeds happen to be found in the codes); iii) the insufficiently documented experimental settings. What works, why and when is still a matter of speculation, although better understanding the limitations of the current state of the art is acknowledged to be a priority. After Ali Rahimi again, *simple experiments, simple theorems are the building blocks that help us understand more complicated systems*. Along this line, [135] propose toy examples to demonstrate and understand the defaults of convergence of gradient descent adversarial learning.
Thirdly, and most importantly, the reported achievements rely on carefully tuned learning architectures and hyper-parameters. The sensitivity of the results to the selection and calibration of algorithms has been identified since the end 80s as a key ML bottleneck, and the field of automatic algorithm selection and calibration, referred to as AutoML or Auto-☆ in the following, is at the ML forefront.

TAU aims to contribute to the ML evolution toward a more mature stage along three dimensions. In the short term, the research done in Auto-☆ will be pursued (section 3.3.1). In the medium term, an information theoretic perspective will be adopted to capture the data structure and to calibrate the learning algorithm *depending on the nature and amount of the available data* (section 3.3.2). In the longer term, our goal is to leverage the methodologies forged in statistical physics to understand and control the trajectories of complex learning systems (section 3.3.3).

*3.3.1. Auto-**

Participants: Isabelle Guyon, Marc Schoenauer, Michèle Sebag
PhD: Guillaume Doquet, Zhengying Liu, Herilalaina Rakotoarison, Lisheng Sun
Collaboration: Olivier Bousquet, André Elisseeff (Google Zurich)

The so-called Auto-☆ task, concerned with selecting a (quasi) optimal algorithm and its hyper-parameters depending on the problem instance at hand, remained a key issue in ML for the last three decades [75], as well as in optimization at large [115], including combinatorial optimization and constraint satisfaction [122], [104] and continuous optimization [71]. This issue, tackled by several European projects along the decades, governs the knowledge transfer to industry, due to the shortage of data scientists. It becomes even more crucial as models are more complex and their training requires more computational resources. This has motivated several international challenges devoted to Auto-ML [113] (see also Section 3.4), including the AutoDL challenge series [129] launched in 2019[0] (see also Section 7.6).

Several approaches have been used to tackle Auto-☆ in the literature, and TAU has been particularly active in several of them. Meta-learning aims to build a surrogate performance model, estimating the performance of an algorithm configuration on *any* problem instance characterized from its meta-feature values [146], [104], [72], [71], [103]. Collaborative filtering, considering that a problem instance "likes better" an algorithm configuration yielding a better performance, learns to recommend good algorithms to problem instances [153], [137]. Bayesian optimization proceeds by alternatively building a surrogate model of algorithm performances on *the* problem instance at hand, and tackling it [95]. This last approach currently is the prominent one; as shown in [137], the meta-features developed for AutoML are hardly relevant, hampering both meta-learning and collaborative filtering. The design of better features is another long-term research direction, in which TAU has recently been [32], ans still is very active. more recent approach used in TAU [40] extends the Bayesian Optimization approach with a Multi-Armed Bandit algorithm to generate the full Machine Learning pipeline, competing with the famed AutoSKLearn [95] (see Section 7.2.1). These results are presented in Section 7.2.1 .

3.3.2. Information theory: adjusting model complexity and data fitting

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag
PhD: Corentin Tallec, Pierre Wolinski, Léonard Blier
Collaboration: Yann Ollivier (Facebook)

In the 60s, Kolmogorov and Solomonoff provided a well-grounded theory for building (probabilistic) models best explaining the available data [147], [108], that is, the shortest programs able to generate these data. Such programs can then be used to generate further data or to answer specific questions (interpreted as missing values in the data). Deep learning, from this viewpoint, efficiently explores a space of computation graphs, described from its hyperparameters (network structure) and parameters (weights). Network training amounts to optimizing these parameters, namely, navigating the space of computational graphs to find a network, as simple as possible, that explain the past observations well.

This vision is at the core of variational auto-encoders [121], directly optimizing a bound on the Kolmogorov complexity of the dataset. More generally variational methods provide quantitative criteria to identify superfluous elements (edges, units) in a neural network, that can potentially be used for structural optimization of the network (Leonard Blier's PhD, started Oct. 2018).

[0] https://autodl.chalearn.org/neurips2019

The same principles apply to unsupervised learning, aimed to find the maximum amount of structure hidden in the data, quantified using this information-theoretic criterion.

The known invariances in the data can be exploited to guide the model design (e.g. as translation invariance leads to convolutional structures, or LSTM is shown to enforce the invariance to time affine transformations of the data sequence [157]). Scattering transforms exploit similar principles [76]. A general theory of how to detect *unknown* invariances in the data, however, is currently lacking.

The view of information theory and Kolmogorov complexity suggests that key program operations (composition, recursivity, use of predefined routines) should intervene when searching for a good computation graph. One possible framework for exploring the space of computation graphs with such operations is that of Genetic Programming. It is interesting to see that evolutionary computation appeared in the last two years among the best candidates to explore the space of deep learning structures [145], [126]. Other approaches might proceed by combining simple models into more powerful ones, e.g. using "Context Tree Weighting" [166] or switch distributions [90]. Another option is to formulate neural architecture design as a reinforcement learning problem [73]; the value of the building blocks (predefined routines) might be defined using e.g., Monte-Carlo Tree Search. A key difficulty is the computational cost of retraining neural nets from scratch upon modifying their architecture; an option might be to use neutral initializations to support warm-restart.

3.3.3. Analyzing and Learning Complex Systems

Participants: Cyril Furtlehner, Aurélien Decelle, François Landes, Michèle Sebag
PhD: Giancarlo Fissore
Collaboration: Enrico Camporeale (CWI); Jacopo Rocchi (LPTMS Paris Sud), the Simons team: Rahul Chako (post-doc), Andrea Liu (UPenn), David Reichman (Columbia), Giulio Biroli (ENS), Olivier Dauchot (ESPCI), Hufei Han (Symantec).

Methods and criteria from statistical physics have been widely used in ML. In early days, the capacity of Hopfield networks (associative memories defined by the attractors of an energy function) was investigated by using the replica formalism [69]. Restricted Boltzmann machines likewise define a generative model built upon an energy function trained from the data. Along the same lines, Variational Auto-Encoders can be interpreted as systems relating the free energy of the distribution, the information about the data and the entropy (the degree of ignorance about the micro-states of the system) [165]. A key promise of the statistical physics perspective and the Bayesian view of deep learning is to harness the tremendous growth of the model size (billions of weights in recent machine translation netwowrks), and make them sustainable through e.g. posterior drop-out [136], weight quantization and probabilistic binary networks [131]. Such "informational cooling" of a trained deep network can reduce its size by several orders of magnitude while preserving its performance.

Statistical physics is among the key expertises of TAU, originally only represented by Cyril Furtlehner, later strenghtened by Aurélien Decelle's and François Landes' arrivals in 2014 and 2018. On-going studies are conducted along several directions.

Generative models are most often expressed in terms of a Gibbs distributions $P[S] = \exp(-E[S])$, where energy E involves a sum of building blocks, modelling the interactions among variables. This formalization makes it natural to use mean-field methods of statistical physics and associated inference algorithms to both train and exploit such models. The difficulty is to find a good trade-off between the richness of the structure and the efficiency of mean-field approaches. One direction of research pursued in TAU, [97] in the context of traffic forecasting, is to account for the presence of cycles in the interaction graph, to adapt inference algorithms to such graphs with cycles, while constraining graphs to remain compatible with mean-field inference.

Another direction, explored in TAO/TAU in the recent years, is based on the definition and exploitation of self-consistency properties, enforcing principled divide-and-conquer resolutions. In the particular case of the message-passing Affinity Propagation algorithm for instance [168], self-consistency imposes the invariance of the solution when handled at different scales, thus enabling to characterize the critical value of the penalty and other hyper-parameters in closed form (in the case of simple data distributions) or empirically otherwise [98].

A more recent research direction examines the quantity of information in a (deep) neural net along the random matrix theory framework [78]. It is addressed in Giancarlo Fissore's PhD, and is detailed in Section 7.2.3 .

Finally, we note the recent surge in using ML to address fundamental physics problems: from turbulence to high-energy physics and soft matter as well (with amorphous materials at its core) [19]. TAU's dual expertise in Deep Networks and in statistical physics places it in an ideal position to significantly contribute to this domain and shape the methods that will be used by the physics community in the future. François Landes' recent arrival in the team makes TAU a unique place for such interdisciplinary research, thanks to his collaborators from the Simons Collaboration *Cracking the Glass Problem* (gathering 13 statistical physics teams at the international level). This project is detailed in Section 7.2.3 .

Independently, François Landes is actively collaborating with statistical physicists (Alberto Rosso, LPTMS, Univ. Paris-Saclay) and physcists at the frontier with geophysics (Eugenio Lippiello, Second Univ. of Naples) [20]. A possible CNRS grant (80Prime) may finance a shared PhD, at the frontier between seismicity and ML (Alberto Rosso, Marc Schoenauer and François Landes).

3.4. Organisation of Challenges

Participants: Cécile Germain, Isabelle Guyon, Marc Schoenauer, Michèle Sebag
Challenges have been an important drive for Machine Learning research for many years, and TAO members have played important roles in the organization of many such challenges: Michèle Sebag was head of the challenge programme in the Pascal European Network of Excellence (2005-2013); Isabelle Guyon, as mentioned, was the PI of many challenges ranging from causation challenges [109], to AutoML [110]. The Higgs challenge [67], most attended ever Kaggle challenge, was jointly organized by TAO (C. Germain), LAL-IN2P3 (D. Rousseau and B. Kegl) and I. Guyon (not yet at TAO), in collaboration with CERN and Imperial College.

TAU was also particularly implicated with the ChaLearn Looking At People (LAP) challenge series in Computer Vision, in collaboration with the University of Barcelona [92] including the Job Candidate Screening Coopetition [91]; the Real Versus Fake Expressed Emotion Challenge (ICCV 2017) [163]; the Large-scale Continuous Gesture Recognition Challenge (ICCV 2017) [163]; the Large-scale Isolated Gesture Recognition Challenge (ICCV 2017) [163].

Other challenges have been organized in 2019, or are planned for the near future, detailed in Section 7.6 . In particular, many of them now run on the Codalab platform, managed by Isabelle Guyon and maintained at LRI.

TOSCA Team

3. Research Program

3.1. Research Program

Most often physicists, economists, biologists and engineers need a stochastic model because they cannot describe the physical, economical, biological, etc., experiment under consideration with deterministic systems, either because of its complexity and/or its dimension or because precise measurements are impossible. Therefore, they abandon trying to get the exact description of the state of the system at future times given its initial conditions, and try instead to get a statistical description of the evolution of the system. For example, they desire to compute occurrence probabilities for critical events such as the overstepping of a given thresholds by financial losses or neuronal electrical potentials, or to compute the mean value of the time of occurrence of interesting events such as the fragmentation to a very small size of a large proportion of a given population of particles. By nature such problems lead to complex modelling issues: one has to choose appropriate stochastic models, which require a thorough knowledge of their qualitative properties, and then one has to calibrate them, which requires specific statistical methods to face the lack of data or the inaccuracy of these data. In addition, having chosen a family of models and computed the desired statistics, one has to evaluate the sensitivity of the results to the unavoidable model specifications. The TOSCA team, in collaboration with specialists of the relevant fields, develops theoretical studies of stochastic models, calibration procedures, and sensitivity analysis methods.

In view of the complexity of the experiments, and thus of the stochastic models, one cannot expect to use closed form solutions of simple equations in order to compute the desired statistics. Often one even has no other representation than the probabilistic definition (e.g., this is the case when one is interested in the quantiles of the probability law of the possible losses of financial portfolios). Consequently the practitioners need Monte Carlo methods combined with simulations of stochastic models. As the models cannot be simulated exactly, they also need approximation methods which can be efficiently used on computers. The TOSCA team develops mathematical studies and numerical experiments in order to determine the global accuracy and the global efficiency of such algorithms.

The simulation of stochastic processes is not motivated by stochastic models only. The stochastic differential calculus allows one to represent solutions of certain deterministic partial differential equations in terms of probability distributions of functionals of appropriate stochastic processes. For example, elliptic and parabolic linear equations are related to classical stochastic differential equations (SDEs), whereas nonlinear equations such as the Burgers and the Navier–Stokes equations are related to McKean stochastic differential equations describing the asymptotic behavior of stochastic particle systems. In view of such probabilistic representations one can get numerical approximations by using discretization methods of the stochastic differential systems under consideration. These methods may be more efficient than deterministic methods when the space dimension of the PDE is large or when the viscosity is small. The TOSCA team develops new probabilistic representations in order to propose probabilistic numerical methods for equations such as conservation law equations, kinetic equations, and nonlinear Fokker–Planck equations.

TRIPOP Project-Team

3. Research Program

3.1. Introduction

In this section, we develop our scientific program. In the framework of nonsmooth dynamical systems, the activities of the project–team will be on focused on the following research axes:

- *Axis 1: Modeling and analysis (detailed in Sect. 3.2).*
- *Axis 2: Numerical methods and simulation (detailed in Sect. 3.3).*
- *Axis 3: Automatic Control (detailed in Sect. 3.4)*

These research axes will be developed with a strong emphasis on the software development and the industrial transfer.

3.2. Axis 1: Modeling and analysis

This axis is dedicated to the modeling and the mathematical analysis of nonsmooth dynamical systems. It consists of four main directions. Two directions are in the continuation of BIPOP activities: 1) multibody vibro-impact systems (Sect. 3.2.1) and 2) excitable systems (Sect. 3.2.2). Two directions are completely new with respect to BIPOP: 3) Nonsmooth geomechanics and natural hazards assessment (Sect. 3.2.3) and 4) Cyber-physical systems (hybrid systems) (Sect. 3.2.4).

3.2.1. Multibody vibro-impact systems

Participants: B. Brogliato, F. Bourrier, G. James, V. Acary

- *Multiple impacts with or without friction :* there are many different approaches to model collisions, especially simultaneous impacts (so-called multiple impacts) [83]. One of our objectives is on one hand to determine the range of application of the models (for instance, when can one use "simplified" rigid contact models relying on kinematic, kinetic or energetic coefficients of restitution?) on typical benchmark examples (chains of aligned beads, rocking block systems). On the other hand, try to take advantage of the new results on nonlinear waves phenomena, to better understand multiple impacts in 2D and 3D granular systems. The study of multiple impacts with (unilateral) nonlinear visco-elastic models (Simon-Hunt-Crossley, Kuwabara-Kono), or visco-elasto-plastic models (assemblies of springs, dashpots and dry friction elements), is also a topic of interest, since these models are widely used.
- *Artificial or manufactured or ordered granular crystals, meta-materials :* Granular metamaterials (or more general nonlinear mechanical metamaterials) offer many perspectives for the passive control of waves originating from impacts or vibrations. The analysis of waves in such systems is delicate due to spatial discreteness, nonlinearity and non-smoothness of contact laws [85], [71], [72], [78]. We will use a variety of approaches, both theoretical (e.g. bifurcation theory, modulation equations) and numerical, in order to describe nonlinear waves in such systems, with special emphasis on energy localization phenomena (excitation of solitary waves, fronts, breathers).
- *Systems with clearances, modeling of friction :* joint clearances in kinematic chains deserve specific analysis, especially concerning friction modeling [38]. Indeed contacts in joints are often conformal, which involve large contact surfaces between bodies. Lubrication models should also be investigated.
- *Painlevé paradoxes :* the goal is to extend the results in [64], which deal with single-contact systems, to multi-contact systems. One central difficulty here is the understanding and the analysis of singularities that may occur in sliding regimes of motion.

As a continuation of the work in the BIPOP team, our software code, Siconos (see Sect. 5.1) will be our favorite software platform for the integration of these new modeling results.

3.2.2. Excitable systems

Participants: A. Tonnelier, G. James

An excitable system elicits a strong response when the applied perturbation is greater than a threshold [80], [81], [42], [90]. This property has been clearly identified in numerous natural and physical systems. In mechanical systems, non-monotonic friction law (of spinodal-type) leads to excitability. Similar behavior may be found in electrical systems such as active compounds of neuristor type. Models of excitable systems incorporate strong non-linearities that can be captured by non-smooth dynamical systems. Two properties are deeply associated with excitable systems: oscillations and propagation of nonlinear waves (autowaves in coupled excitable systems). We aim at understanding these two dynamical states in excitable systems through theoretical analysis and numerical simulations. Specifically we plan to study:

- Threshold-like models in biology: spiking neurons, gene networks.
- Frictional contact oscillators (slider block, Burridge-Knopoff model).
- Dynamics of active electrical devices : memristors, neuristors.

3.2.3. Nonsmooth geomechanics and natural hazards assessment

Participants: F. Bourrier, B. Brogliato, G. James, V. Acary

- *Rockfall impact modeling :* Trajectory analysis of falling rocks during rockfall events is limited by a rough modeling of the impact phase [44], [43], [76]. The goal of this work is to better understand the link between local impact laws at contact with refined geometries and the efficient impact laws written for a point mass with a full reset map. A continuum of models in terms of accuracy and complexity will be also developed for the trajectory studies. In particular, nonsmooth models of rolling friction, or rolling resistance will be developed and formulated using optimization problems.
- *Experimental validation :* The participation of IRSTEA with F. Bourrier makes possible the experimental validation of models and simulations through comparisons with real data. IRSTEA has a large experience of lab and in-situ experiments for rockfall trajectories modeling [44], [43]. It is a unique opportunity to vstrengthen our model and to prove that nonsmooth modeling of impacts is reliable for such experiments and forecast of natural hazards.
- *Rock fracturing :* When a rock falls from a steep cliff, it stores a large amount of kinetic energy that is partly dissipated though the impact with the ground. If the ground is composed of rocks and the kinetic energy is sufficiently high, the probability of the fracture of the rock is high and yields an extra amount of dissipated energy but also an increase of the number of blocks that fall. In this item, we want to use the capability of the nonsmooth dynamical framework for modeling cohesion and fracture [73], [36] to propose new impact models.
- *Rock/forest interaction :* To prevent damages and incidents to infrastructures, a smart use of the forest is one of the ways to control trajectories (decrease of the run-out distance, jump heights and the energy) of the rocks that fall under gravity [56], [58]. From the modeling point of view and to be able to improve the protective function of the forest, an accurate modeling of impacts between rocks and trees is required. Due to the aspect ratio of the trees, they must be considered as flexible bodies that may be damaged by the impact. This new aspect offers interesting modeling research perspectives.

More generally, our collaboration with IRSTEA opens new long term perspectives on granular flows applications such as debris and mud flows, granular avalanches and the design of structural protections. *The numerical methods that go with these new modeling approaches will be implemented in our software code, Siconos (see Sect. 5.1)*

3.2.4. Cyber-physical systems (hybrid systems)

Participants: V. Acary, B. Brogliato, C. Prieur, A. Tonnelier

Nonsmooth systems have a non-empty intersection with hybrid systems and cyber–physical systems. However, nonsmooth systems enjoy strong mathematical properties (concept of solutions, existence and uniqueness) and efficient numerical tools. This is often the result of the fact that nonsmooth dynamical systems are models of physical systems, and then, take advantage of their intrinsic property (conservation or dissipation of energy, passivity, stability). A standard example is a circuit with n ideal diodes. From the hybrid point of view, this circuit is a piecewise smooth dynamical system with 2^n modes, that can be quite cumbersome to enumerate in order to determinate the current mode. As a nonsmooth system, this circuit can be formulated as a complementarity system for which there exist efficient time–stepping schemes and polynomial time algorithms for the computation of the current mode. The key idea of this research action is to take benefit of this observation to improve the hybrid system modeling tools.

Research actions: There are two main actions in this research direction that will be implemented in the framework of the Inria Project Lab (IPL " Modeliscale", see https://team.inria.fr/modeliscale/ for partners and details of the research program):

•*Structural analysis of multimode DAE :* When a hybrid system is described by a Differential Algebraic Equation (DAE) with different differential indices in each continuous mode, the structural analysis has to be completely rethought. In particular, the re-initialization rule, when a switching occurs from a mode to another one, has to be consistently designed. We propose in this action to use our knowledge in complementarity and (distribution) differential inclusions [30] to design consistent re-initialization rule for systems with nonuniform relative degree vector $(r_1, r_2, ..., r_m)$ and $r_i \neq r_j, i \neq j$.

•*Cyber–physical in hybrid systems modeling languages :* Nowadays, some hybrid modeling languages and tools are widely used to describe and to simulate hybrid systems (MODELICA, SIMULINK, and see [53] for references therein). Nevertheless, the compilers and the simulation engines behind these languages and tools suffer from several serious weaknesses (failure, weird output or huge sensitivity to simulation parameters), especially when some components, that are standard in nonsmooth dynamics, are introduced (piecewise smooth characteristic, unilateral constraints and complementarity condition, relay characteristic, saturation, dead zone, ...). One of the main reasons is the fact that most of the compilers reduce the hybrid system to a set of smooth modes modeled by differential algebraic equations and some guards and reinitialization rules between these modes. Sliding mode and Zeno–behaviour are really harsh for hybrid systems and relatively simple for nonsmooth systems. With B. Caillaud (Inria HYCOMES) and M. Pouzet (Inria PARKAS), we propose to improve this situation by implementing a module able to identify/describe nonsmooth elements and to efficiently handle them with SICONOS as the simulation engine. They have already carried out a first implementation [51] in Zelus, a synchronous language for hybrid systems http://zelus.di.ens.fr. Removing the weaknesses related to the nonsmoothness of solutions should improve hybrid systems towards robustness and certification.

•*A general solver for piecewise smooth systems* This direction is the continuation of the promising result on modeling and the simulation of piecewise smooth systems [35]. As for general hybrid automata, the notion or concept of solutions is not rigorously defined from the mathematical point of view. For piecewise smooth systems, multiplicity of solutions can happen and sliding solutions are common. The objective is to recast general piecewise smooth systems in the framework of differential inclusions with Aizerman–Pyatnitskii extension [35], [60]. This operation provides a precise meaning to the concept of solutions. Starting from this point, the goal is to design and study an efficient numerical solver (time–integration scheme and optimization solver) based on an equivalent formulation as mixed complementarity systems of differential variational inequalities. We are currently discussing the issues in the mathematical analysis. The goal is to prove the convergence of the time–stepping scheme to get an existence theorem. With this work, we should also be able to discuss the general Lyapunov stability of stationary points of piecewise smooth systems.

3.3. Axis 2: Numerical methods and simulation

This axis is dedicated to the numerical methods and simulation for nonsmooth dynamical systems. As we mentioned in the introduction, the standard numerical methods have been largely improved in terms of accuracy and dissipation properties in the last decade. Nevertheless, the question of the geometric

time–integration techniques remains largely open. It constitutes the objective of the first research direction in Sect. 3.3.1 . Beside the standard IVP, the question of normal mode analysis for nonsmooth systems is also a research topic that emerged in the recent years. More generally, the goal of the second research direction (Sect. 3.3.2) is to develop numerical methods to solve boundary value problems in the nonsmooth framework. This will serve as a basis for the computation of the stability and numerical continuation of invariants. Finally, once the time-integration method is chosen, it remains to solve the one-step nonsmooth problem, which is, most of time, a numerical optimization problem. In Sect. 3.3.3 , we propose to study two specific problems with a lot of applications: the Mathematical Program with Equilibrium Constraints (MPEC) for optimal control, and Second Order Cone Complementarity Problems (SOCCP) for discrete frictional contact systems. After some possible prototypes in scripting languages (Python and Matlab), we will be attentive that all these developments of numerical methods will be integrated in Siconos.

3.3.1. Geometric time–integration schemes for nonsmooth Initial Value Problem (IVP)

Participants: V. Acary, B. Brogliato, G. James, F. Pérignon

The objective of this research item is to continue to improve classical time–stepping schemes for nonsmooth systems to ensure some qualitative properties in discrete-time. In particular, the following points will be developed

- Conservative and dissipative systems. The question of the energy conservation and the preservation of dissipativity properties in the Willems sense [63] will be pursued and extended to new kinds of systems (nonlinear mechanical systems with nonlinear potential energy, systems with limited differentiability (rigid impacts vs. compliant models)).
- Lie–group integration schemes for finite rotations for the multi-body systems extending recent progresses in that directions for smooth systems [40].
- Conservation and preservation of the dispersion properties of the (non)-dispersive system.

3.3.2. Stability and numerical continuation of invariants

Participants: G. James, V. Acary, A. Tonnelier, F. Pérignon,

By invariants, we mean equilibria, periodic solutions, limit cycles or waves. Our preliminary work on this subject raised the following research perspectives:

- Computation of periodic solutions of discrete mechanical systems . The modal analysis, *i.e.*, a spectral decomposition of the problem into linear normal modes is one of the basic tools for mechanical engineers to study dynamic response and resonance phenomena of an elastic structure. Since several years, the concept of nonlinear normal modes [74], that is closely related to the computation of quasi-periodic solutions that live in a nonlinear manifold, has emerged as the nonlinear extension of the modal analysis. One of the fundamental question is: what remains valid if we add unilateral contact conditions ? The computation of nonsmooth modes amounts to computing periodic solutions, performing the parametric continuation of solution branches and studying the stability of these branches. This calls for time integration schemes for IVP an BVP that satisfy some geometric criteria: conservation of energy, reduced numerical dispersion, symplecticity as we described before. Though the question of conservation of energy for unilateral contact has been discussed in [25], the other questions remain open. For the shooting technique and the study of stability, we need to compute the Jacobian matrix of the flow with respect to initial conditions, the so-called saltation matrix [75], [84] for nonsmooth flows. The eigenvalues of this matrix are the Floquet multipliers that give some information on the stability of the periodic solutions. The question of an efficient computation of this matrix is also an open question. For the continuation, the question is also largely open since the continuity of the solutions with respect to the parameters is not ensured.
- Extension to elastic continuum media . This is a difficult task. First of all, the question of the mathematical model for the dynamic continuum problem with unilateral contact raises some problems of well–posedness. For instance, the need for an impact law is not clear in some cases. If we perform

a semi–discretization in space with classical techniques (Finite Element Methods, Finite Difference Schemes), we obtain a discrete system for which the impact law is needed. Besides all the difficulties that we enumerate for discrete systems in the previous paragraph, the space discretization also induces numerical dispersion that may destroy the periodic solutions or renders their computation difficult. The main targeted applications for this research are cable–systems, string musical instruments, and seismic response of electrical circuit breakers with Schneider Electric.

- Computation of solutions of nonsmooth time Boundary Value Problems (BVP) (collocation, shooting) . The technique developed in the two previous items can serve as a basis for the development of more general solvers for nonsmooth BVP that can be for instance found when we solve optimal control problems by direct or indirect methods, or the computation of nonlinear waves. Two directions can be envisaged:
  - Shooting and multiple shooting techniques. In such methods, we reformulate the BVP into a sequence of IVPs that are iterated through a Newton based technique. This implies the computation of Jacobians for nonsmooth flows, the question of the continuity w.r.t to initial condition and the use of semi-smooth Newton methods.
  - Finite differences and collocations techniques. In such methods, the discretization will result into a large sparse optimization problems to solve. The open questions are as follows: a) the study of convergence, b) how to locally improve the order if the solution is locally smooth, and c) how to take benefit of spectral methods.
- Continuation techniques of solutions with respect to a parameter. Standard continuation technique requires smoothness. What types of methods can be extended in the nonsmooth case (arc-length technique, nonsmooth (semi-smooth) Newton, Asymptotical Numerical Methods (ANM))

3.3.3. Numerical optimization for discrete nonsmooth problems

Participants: V. Acary, M. Brémond, F. Pérignon, B. Brogliato, C. Prieur

- Mathematical Program with Equilibrium Constraints (MPEC) for optimal control . The discrete problem that arises in nonsmooth optimal control is generally a MPEC [91]. This problem is intrinsically nonconvex and potentially nonsmooth. Its study from a theoretical point of view has started 10 years ago but there is no consensus for its numerical solving. The goal is to work with world experts of this problem (in particular M. Ferris from Wisconsin University) to develop dedicated algorithms for solving MPEC, and provide to the optimization community challenging problems.
- Second Order Cone Complementarity Problems (SOCCP) for discrete frictional systems : After some extensive comparisons of existing solvers on a large collection of examples [33], [27], the numerical treatment of constraints redundancy by the proximal point technique and the augmented Lagrangian formulation seems to be a promising path for designing new methods. From the comparison results, it appears that the redundancy of constraints prevents the use of second order methods such as semi–smooth Newton methods or interior point methods. With P. Armand (XLIM, U. de Limoges), we propose to adapt recent advances for regularizing constraints for the quadratic problem [61] for the second-order cone complementarity problem. The other question is the improvement of the efficiency of the algorithms by using accelerated schemes for the proximal gradient method that come from large–scale machine learning and image processing problems. Learning from the experience in large–scale machine learning and image processing problems, the accelerated version of the classical gradient algorithm [82] and the proximal point algorithm [41], and many of their further extensions, could be of interest for solving discrete frictional contact problems. Following the visit of Y. Kanno (University of Tokyo) and his preliminary experience on frictionless problems, we will extend its use to frictional contact problem. When we face large-scale problems, the main available solvers is based on a Gauss–Seidel strategy that is intrinsically sequential. Accelerated first-order methods could be a good alternative to take benefit of the distributed scientific computing architectures.

3.4. Axis 3: Automatic Control

Participants: B. Brogliato, C. Prieur, V. Acary

This last axis is dedicated to the automatic control of nonsmooth dynamical systems, or the nonsmooth control of smooth systems. The first item concerns the discrete-time sliding mode control for which significant results on the implicit implementation have been obtained in the BIPOP team. The idea is to pursue this research towards state observers and differentiators (Sect 3.4.1). The second direction concerns the optimal control which brings of nonsmoothness in their solution and their formulation. After the preliminary work in BIPOP on the quadratic optimal control of Linear Complementarity systems(LCS), we propose to go further to the minimal time problem, to impacting systems and optimal control with state constraints (Sect. 3.4.2). In Sect 3.4.3 , the objective is to study the control of nonsmooth systems that contain unilateral constraint, impact and friction. The targeted systems are cable–driven systems, multi-body systems with clearances and granular materials. In Sect 3.4.4 , we will continue our work on the higher order Moreau sweeping process. Up to now, the work of BIPOP was restricted to finite-dimensional systems. In Sect 3.4.5 , we propose to extend our approach to the control of elastic structures subjected to contact unilateral constraints.

It is noteworthy that most of the problems listed below, will make strong use of the numerical tools analyzed in Axis 2, and of the Modeling analysis of Axis 1. For instance all optimal control problems yield BVPs. Control of granular materials will undoubtedly use models and numerical simulation developed in Axis 1 and 2. And so on. It has to be stressed that the type of nonsmooth models we are working with, deserve specific numerical algorithms which cannot be found in commercial software packages. One of the goals is to continue to extend our software package Siconos, and in particular the siconos/control toolbox with these developments.

3.4.1. Discrete-time Sliding-Mode Control (SMC) and State Observers (SMSO)

- *SMSO, exact differentiators:* we have introduced and obtained significant results on the implicit discretization of various classes of sliding-mode controllers [29], [31], [68], [79], [47], with successful experimental validations [69], [68], [70], [92]. Our objective is to prove that the implicit discretization can also bring advantages for sliding-mode state observers and Levant's exact differentiators, compared with the usual explicit digital implementation that generates chattering. In particular the implicit discretization guarantees Lyapunov stability and finite-time convergence properties which are absent in explicit methods.
- *High-Order SMC (HOSMC):* this family of controllers has become quite popular in the sliding-mode scientific community since its introduction by Levant in the nineties. We want here to continue the study of implicit discretization of HOSMC (twisting, super-twisting algorithms) and especially we would like to investigate the comparisons between classical (first order) SMC and HOSMC, when both are implicitly discretized, in terms of performance, accuracy, chattering suppression. Another topic of interest is stabilization in finite-time of systems with impacts and unilateral constraints, in a discrete-time setting.

3.4.2. Optimal Control

- *Linear Complementarity Systems (LCS) :* With the PhD thesis of A. Vieira, we have started to study the quadratic optimal control of LCS. Our objective is to go further with minimum-time problems. Applications of LCS are mainly in electrical circuits with set-valued components such as ideal diodes, transistors, *etc*. Such problems naturally yield MPEC when numerical solvers are sought. It is therefore intimately linked with Axis 2 objectives.
- *Impacting systems :* the optimal control of mechanical systems with unilateral constraints and impacts, largely remains an open issue. The problem can be tackled from various approaches: vibro-impact systems (no persistent contact modes) that may be transformed into discrete-time mappings *via* the impact Poincaré map; or the classical integral action minimization (Bolza problem) subjected to the complementarity Lagrangian dynamics including impacts.

- *State constraints, generalized control :* this problem differs from the previous two, since it yields Pontryagin's first order necessary conditions that take the form of an LCS with higher relative degree between the complementarity variables. This is related to the numerical techniques for the higher order sweeping process [30].

3.4.3. Control of nonsmooth discrete Lagrangian systems

- *Cable–driven systems:* these systems are typically different from the cable-car systems, and are closer in their mechanical structure to so-called tensegrity structures. The objective is to actuate a system *via* cables supposed in a first instance to be flexible (slack mode) but non-extensible in their longitudinal direction. This gives rise to complementarity conditions, one big difference with usual complementarity Lagrangian systems being that the control actions operate directly in one of the complementary variables (and not in the smooth dynamics as in cable-car systems). Therefore both the cable models and the control properties are expected to differ a lot from what we may use for cableway systems (for which guaranteeing a positive cable tension is usually not an issue, hence avoiding slack modes, but the deformation of the cables due to the nacelles and cables weights, is an important factor). Tethered systems are a close topic.
- *Multi-body systems with clearances:* our approach is to use models of clearances with dynamical impact effects, *i.e.* within Lagrangian complementarity systems. Such systems are strongly underactuated due to mechanical play at the joints. However their structure, as underactuated systems, is quite different from what has been usually considered in the Robotics and Control literature. In the recent past we have proposed a thorough numerical robustness analysis of various feedback collocated and non-collocated controllers (PD, linearization, passivity-based). We propose here to investigate specific control strategies tailored to such underactuated systems [46].
- *Granular systems:* the context is the feedback control of granular materials. To fix the ideas, one may think of a "juggling" system whose "object" (uncontrolled) part consists of a chain of aligned beads. Once the modeling step has been fixed (choice of a suitable multiple impact law), one has to determine the output to be controlled: all the beads, some of the beads, the chain's center of mass (position, velocity, vibrational magnitude and frequency), *etc*. Then we aim at investigating which type of controller may be used (output or state feedback, "classical" or sinusoidal input with feedback through the magnitude and frequency) and especially which variables may be measured/observed (positions and/or velocities of all or some of the beads, position and/or velocity of the chain's center of gravity). This topic follows previous results we obtained on the control of juggling systems [48], with increasing complexity of the "object"'s dynamics. The next step would be to extend to 2D and then 3D granular materials. Applications concern vibrators, screening, transport in mining and manufacturing processes.
- *Stability of structures:* our objective here is to study the stability of stacked blocks in 2D or 3D, and the influence on the observed behavior (numerically and/or analytically) of the contact/impact model.

3.4.4. Switching LCS and DAEs, higher-order sweeping process (HOSwP)

- We have gained a strong experience in the field of complementarity systems and distribution differential inclusions [30], [49], that may be seen as some kind of switching DAEs. We plan to go further with non-autonomous HOSwP with switching feedback inputs and non-uniform vector relative degrees. Switching linear complementarity systems can also be studied, though the exact relationships between both point of views remain unclear at the present time. This axis of research is closely related to cyber-physical systems in section 3.2 .

3.4.5. Control of Elastic (Visco-plastic) systems with contact, impact and friction

- *Stabilization, trajectory tracking:* until now we have focused on the stability and the feedback control of systems of rigid bodies. The proposal here is to study the stabilization of flexible systems (for instance, a "simple" beam) subjected to unilateral contacts with or without set-valued friction

(contacts with obstacles, or impacts with external objects line particle/beam impacts). This gives rise to varying (in time and space) boundary conditions. The best choice of a good contact law is a hard topic discussed in the literature.

- *Cableway systems (STRMTG, POMA):* cable-car systems present challenging control problems because they usually are underactuated systems, with large flexibilities and deformations. Simplified models of cables should be used (Ritz-Galerkin approach), and two main classes of systems may be considered: those with moving cable and only actuator at the station, and those with fixed cable but actuated nacelles. It is expected that they possess quite different control properties and thus deserve separate studies. The nonsmoothness arises mainly from the passage of the nacelles on the pylons, which induces frictional effects and impacts. It may certainly be considered as a nonsmooth set-valued disturbance within the overall control problem.

TROPICAL Project-Team

3. Research Program

3.1. Optimal control and zero-sum games

The dynamic programming approach allows one to analyze one or two-player dynamic decision problems by means of operators, or partial differential equations (Hamilton–Jacobi or Isaacs PDEs), describing the time evolution of the value function, i.e., of the optimal reward of one player, thought of as a function of the initial state and of the horizon. We work especially with problems having long or infinite horizon, modelled by stopping problems, or ergodic problems in which one optimizes a mean payoff per time unit. The determination of optimal strategies reduces to solving nonlinear fixed point equations, which are obtained either directly from discrete models, or after a discretization of a PDE.

The geometry of solutions of optimal control and game problems Basic questions include, especially for stationary or ergodic problems, the understanding of existence and uniqueness conditions for the solutions of dynamic programming equations, for instance in terms of controllability or ergodicity properties, and more generally the understanding of the structure of the full set of solutions of stationary Hamilton–Jacobi PDEs and of the set of optimal strategies. These issues are already challenging in the one-player deterministic case, which is an application of choice of tropical methods, since the Lax-Oleinik semigroup, i.e., the evolution semigroup of the Hamilton-Jacobi PDE, is a linear operator in the tropical sense. Recent progress in the deterministic case has been made by combining dynamical systems and PDE techniques (weak KAM theory [72]), and also using metric geometry ideas (abstract boundaries can be used to represent the sets of solutions [86], [4]). The two player case is challenging, owing to the lack of compactness of the analogue of the Lax-Oleinik semigroup and to a richer geometry. The conditions of solvability of ergodic problems for games (for instance, solvability of ergodic Isaacs PDEs), and the representation of solutions are only understood in special cases, for instance in the finite state space case, through tropical geometry and non-linear Perron-Frobenius methods [38], [41], [3].

Algorithmic aspects: from combinatorial algorithms to the attenuation of the curse of dimensionality Our general goal is to push the limits of solvable models by means of fast algorithms adapted to large scale instances. Such instances arise from discrete problems, in which the state space may so large that it is only accessible through local oracles (for instance, in some web ranking applications, the number of states may be the number of web pages) [73]. They also arise from the discretization of PDEs, in which the number of states grows exponentially with the number of degrees of freedom, according to the "curse of dimensionality". A first line of research is the development of *new approximation methods for the value function*. So far, classical approximations by linear combinations have been used, as well as approximation by suprema of linear or quadratic forms, which have been introduced in the setting of dual dynamic programming and of the so called "max-plus basis methods" [74]. We believe that more concise or more accurate approximations may be obtained by unifying these methods. Also, some max-plus basis methods have been shown to *attenuate the curse of dimensionality* for very special problems (for instance involving switching) [97], [78]. This suggests that the complexity of control or games problems may be measured by more subtle quantities that the mere number of states, for instance, by some forms of metric entropy (for example, certain large scale problems have a low complexity owing to the presence of decomposition properties, "highway hierarchies", etc.). A second line of of our research is the development of *combinatorial algorithms*, to solve large scale zero-sum two-player problems with discrete state space. This is related to current open problems in algorithmic game theory. In particular, the existence of polynomial-time algorithms for games with ergodic payment is an open question. See e.g. [43] for a polynomial time average complexity result derived by tropical methods. The two lines of research are related, as the understanding of the geometry of solutions allows to develop better approximation or combinatorial algorithms.

3.2. Non-linear Perron-Frobenius theory, nonexpansive mappings and metric geometry

Several applications (including population dynamics [10] and discrete event systems [56], [64], [46]) lead to studying classes of dynamical systems with remarkable properties: preserving a cone, preserving an order, or being nonexpansive in a metric. These can be studied by techniques of non-linear Perron-Frobenius theory [3] or metric geometry [11]. Basic issues concern the existence and computation of the "escape rate" (which determines the throughput, the growth rate of the population), the characterizations of stationary regimes (non-linear fixed points), or the study of the dynamical properties (convergence to periodic orbits). Nonexpansive mappings also play a key role in the "operator approach" to zero-sum games, since the one-day operators of games are nonexpansive in several metrics, see [8].

3.3. Tropical algebra and convex geometry

The different applications mentioned in the other sections lead us to develop some basic research on tropical algebraic structures and in convex and discrete geometry, looking at objects or problems with a "piecewise-linear " structure. These include the geometry and algorithmics of tropical convex sets [49], [40], tropical semialgebraic sets [52], the study of semi-modules (analogues of vector spaces when the base field is replaced by a semi-field), the study of systems of equations linear in the tropical sense, investigating for instance the analogues of the notions of rank, the analogue of the eigenproblems [42], and more generally of systems of tropical polynomial equations. Our research also builds on, and concern, classical convex and discrete geometry methods.

3.4. Tropical methods applied to optimization, perturbation theory and matrix analysis

Tropical algebraic objects appear as a deformation of classical objects thought various asymptotic procedures. A familiar example is the rule of asymptotic calculus,

$$e^{-a/\epsilon} + e^{-b/\epsilon} \asymp e^{-\min(a,b)/\epsilon} \ , \qquad e^{-a/\epsilon} \times e^{-b/\epsilon} = e^{-(a+b)/\epsilon} \ , \tag{55}$$

when $\epsilon \to 0^+$. Deformations of this kind have been studied in different contexts: large deviations, zero-temperature limits, Maslov's "dequantization method" [96], non-archimedean valuations, log-limit sets and Viro's patchworking method [122], etc.

This entails a relation between classical algorithmic problems and tropical algorithmic problems, one may first solve the $\epsilon = 0$ case (non-archimedean problem), which is sometimes easier, and then use the information gotten in this way to solve the $\epsilon = 1$ (archimedean) case.

In particular, tropicalization establishes a connection between polynomial systems and piecewise affine systems that are somehow similar to the ones arising in game problems. It allows one to transfer results from the world of combinatorics to "classical" equations solving. We investigate the consequences of this correspondence on complexity and numerical issues. For instance, combinatorial problems can be solved in a robust way. Hence, situations in which the tropicalization is faithful lead to improved algorithms for classical problems. In particular, scalings for the polynomial eigenproblems based on tropical preprocessings have started to be used in matrix analysis [80], [84].

Moreover, the tropical approach has been recently applied to construct examples of linear programs in which the central path has an unexpectedly high total curvature [44], and it has also led to positive polynomial-time average case results concerning the complexity of mean payoff games. Similarly, we are studying semidefinite programming over non-archimedean fields [52], [51], with the goal to better understand complexity issues in classical semidefinite and semi-algebraic programming.

VALSE Project-Team

3. Research Program

3.1. Research Program

Valse team works in the domains of control science: dynamical systems, stability analysis, estimation and automatic control. *Our developments are focused on the theoretical and applied aspects related to control and estimation of large-scale multi-sensor and multi-actuator systems based on the use of the theories of finite-time/fixed-time/hyperexponential convergence and homogeneous systems*. The Lyapunov function method and other methods of analysis of dynamical systems form a basis for the studies in Valse team.

The key idea of research program for the team is that a fast (non-asymptotic) convergence of the regulation and estimation errors increases the reliability of intelligent distributed actuators and sensors in complex scenarios, such as interconnected cyber-physical systems (CPSs).

The expertise of Valse's members in theoretical developments of control and estimation theory (finite-time control and estimation algorithms in centralized context [84], [70], [81], [80], [77], homogeneity framework for differential equations [85], [72], [71], [73], [75], [86], [82], time-delay systems [74], [76], [89], distributed systems [83] and algebraic-based methods for estimation [87], [88]) is an essential ingredient to achieve our objective.

The generic chart of different goals and tasks included in the scientific work program of Valse, and interrelations between them, are presented in Fig. 1 . We have selected three main objectives to pursuit with the related tasks to fulfill:

- The first objective consists in design of control and estimation solutions for CPS and IoT, which is the principal aim of Valse, it will contain the main outcomes of our research.
- The second objective is more theoretical, which is needed to make the basement for our design and analysis parts in the previous goal.
- The third objective deals with applications, which will drive the team and motivate the theoretical studies and selected design performances.

All these objectives are interconnected: from a particular problem in an IoT application, it is planned to design a control or estimation algorithm, which leads to development of theoretical tools; and *vice versa*, a new theoretical advance can provide a possibility for development of novel tools, which can be used in applications.

To explain our motivation: *why to use finite-time?* Applying any method for control/estimation has a price in terms of its advantages and disadvantages. There is no universal framework that is the best always and everywhere. Finite-time may appear as a luxurious property for a physical system, requiring the use of nonlinear tools. Of course, if an asymptotic convergence and a linear model are enough for solving a given problem, then there is no reason to develop something else. However, most of the present problems in CPS and IoT are nonlinear (i.e. they have various local behaviors that cannot be collected in only one linear model). Design and analysis of various local linearized models and solutions are luxurious, too. The theory of homogeneity can go beyond linearity offering many new features, while not appearing as severe as other nonlinear tools and having almost all hints of the linear framework. Suppose that, thanks to the homogeneity theory, finite-time/fixed-time can be obtained with a limited difficulty, while adding the bonuses of a stronger robustness and a faster convergence compared to the linear case? *We are convinced that the price of going beyond linear control and estimation can be strongly dropped down by maturing the theory of homogeneity and finite/fixed-time convergence. And also, convinced that it will be compensated in terms of robustness and speed, which can be demanded in the new areas of application as IoT, for example*.

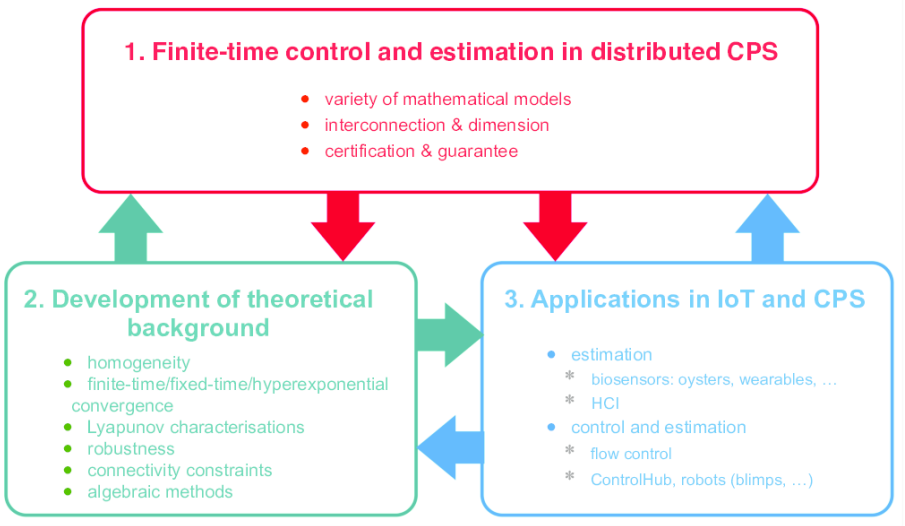


Figure 1. Structure of the objectives and tasks treated in Valse

ABS Project-Team

3. Research Program

3.1. Introduction

The research conducted by ABS focuses on three main directions in Computational Structural Biology (CSB), together with the associated methodological developments:
– Modeling interfaces and contacts,
– Modeling macro-molecular assemblies,
– Modeling the flexibility of macro-molecules,
– Algorithmic foundations.

3.2. Modeling interfaces and contacts

Keywords: Docking, interfaces, protein complexes, structural alphabets, scoring functions, Voronoi diagrams, arrangements of balls.

The Protein Data Bank, http://www.rcsb.org/pdb, contains the structural data which have been resolved experimentally. Most of the entries of the PDB feature isolated proteins [0], the remaining ones being protein - protein or protein - drug complexes. These structures feature what Nature does – up to the bias imposed by the experimental conditions inherent to structure elucidation, and are of special interest to investigate non-covalent contacts in biological complexes. More precisely, given two proteins defining a complex, interface atoms are defined as the atoms of one protein *interacting* with atoms of the second one. Understanding the structure of interfaces is central to understand biological complexes and thus the function of biological molecules [44]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [47]. Current investigations follow two routes. From the experimental perspective [30], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [41]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [36].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [0], or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [25], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i– to be located at distance r, the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [45], [32]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i/q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [37]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

Describing interfaces poses problems in two settings: static and dynamic.

[0] For structures resolved by crystallography, the PDB contains the asymmetric unit of the crystal. Determining the biological unit from the asymmetric unit is a problem in itself.

[0] The Gibbs free energy of a system is defined by $G = H - TS$, with $H = U + PV$. G is minimum at an equilibrium, and differences in G drive chemical reactions.

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as to relate the geometry and the bio-chemistry at the interface level [12]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [26]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [46], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [29].

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [40]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling macro-molecular assemblies

Keywords: Macro-molecular assembly, reconstruction by data integration, proteomics, modeling with uncertainties, curved Voronoi diagrams, topological persistence.

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [24]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [23], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with Uncertainties and Model Assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [22], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [22]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the flexibility of macro-molecules

Keywords: Folding, docking, energy landscapes, induced fit, molecular dynamics, conformers, conformer ensembles, point clouds, reconstruction, shape learning, Morse theory.

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the *free energy* of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called *conformers*, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed [0]. Understanding correlations is of special interest to predict the folding pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

Parameterizing these correlated motions, describing the corresponding energy landscapes, as well as handling collections of conformations pose challenging algorithmic problems.

At the side-chain level, the question of improving rotamer libraries is still of interest [28]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [43]. The standard methodology to analyze such a point cloud consists of resorting to normal modes. Recently, though, more elaborate methods resorting to more local analysis [39], to Morse theory [34] and to analysis of meta-stable states of time series [35] have been proposed.

3.5. Algorithmic foundations

Keywords: Computational geometry, computational topology, optimization, data analysis.

Making a stride towards a better understanding of the biophysical questions discussed in the previous sections requires various methodological developments, which we briefly discuss now.

3.5.1. Modeling Interfaces and Contacts

In modeling interfaces and contacts, one may favor geometric or topological information.

On the geometric side, the problem of modeling contacts at the atomic level is tantamount to encoding multi-body relations between an atom and its neighbors. On the one hand, one may use an encoding of neighborhoods based on geometric constructions such as Voronoi diagrams (affine or curved) or arrangements of balls. On the other hand, one may resort to clustering strategies in higher dimensional spaces, as the p neighbors of a given atom are represented by $3p - 6$ degrees of freedom – the neighborhood being invariant upon rigid motions. The information gathered while modeling contacts can further be integrated into interface models.

On the topological side, one may favor constructions which remain stable if each atom in a structure *retains* the same neighbors, even though the 3D positions of these neighbors change to some extent. This process is observed in flexible docking cases, and call for the development of methods to encode and compare shapes undergoing tame geometric deformations.

3.5.2. Modeling Macro-molecular Assemblies

In dealing with large assemblies, a number of methodological developments are called for.

[0] Assuming local forces are prominent, which in turn subsumes electrostatic interactions are not prominent.

On the experimental side, of particular interest is the disambiguation of proteomics signals. For example, TAP and mass spectrometry data call for the development of combinatorial algorithms aiming at unraveling pairwise contacts between proteins within an assembly. Likewise, density maps coming from electron microscopy, which are often of intermediate resolution (5-10Å) call the development of noise resilient segmentation and interpretation algorithms. The results produced by such algorithms can further be used to guide the docking of high resolutions crystal structures into maps.

As for modeling, two classes of developments are particularly stimulating. The first one is concerned with the design of algorithms performing reconstruction by data integration, a process reminiscent from non convex optimization. The second one encompasses assessment methods, in order to single out the reconstructions which best comply with the experimental data. For that endeavor, the development of geometric and topological models accommodating uncertainties is particularly important.

3.5.3. Modeling the Flexibility of Macro-molecules

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describe the topography of the energy landscape (a question reminiscent from Morse theory)? Can one infer the effective number of degrees of freedom of the system over the simulation, and is this number varying? Answers to these questions would be of major interest to refine our understanding of folding and docking, with applications to the prediction of structural properties. It should be noted in passing that such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [38].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples – the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison; the study of Morse-like constructions stemming from distance functions to points; the analysis of topological invariants of the model and the samples, and their comparison.

AIRSEA Project-Team

3. Research Program

3.1. Introduction

Recent events have raised questions regarding the social and economic implications of anthropic alterations of the Earth system, i.e. climate change and the associated risks of increasing extreme events. Ocean and atmosphere, coupled with other components (continent and ice) are the building blocks of the Earth system. A better understanding of the ocean atmosphere system is a key ingredient for improving prediction of such events. Numerical models are essential tools to understand processes, and simulate and forecast events at various space and time scales. Geophysical flows generally have a number of characteristics that make it difficult to model them. This justifies the development of specifically adapted mathematical methods:

- Geophysical flows are strongly non-linear. Therefore, they exhibit interactions between different scales, and unresolved small scales (smaller than mesh size) of the flows have to be **parameterized** in the equations.
- Geophysical fluids are non closed systems. They are open-ended in their scope for including and dynamically coupling different physical processes (e.g., atmosphere, ocean, continental water, etc). **Coupling** algorithms are thus of primary importance to account for potentially significant feedback.
- Numerical models contain parameters which cannot be estimated accurately either because they are difficult to measure or because they represent some poorly known subgrid phenomena. There is thus a need for **dealing with uncertainties**. This is further complicated by the turbulent nature of geophysical fluids.
- The computational cost of geophysical flow simulations is huge, thus requiring the use of **reduced models, multiscale methods** and the design of algorithms ready for **high performance computing** platforms.

Our scientific objectives are divided into four major points. The first objective focuses on developing advanced mathematical methods for both the ocean and atmosphere, and the coupling of these two components. The second objective is to investigate the derivation and use of model reduction to face problems associated with the numerical cost of our applications. The third objective is directed toward the management of uncertainty in numerical simulations. The last objective deals with efficient numerical algorithms for new computing platforms. As mentioned above, the targeted applications cover oceanic and atmospheric modeling and related extreme events using a hierarchy of models of increasing complexity.

3.2. Modeling for oceanic and atmospheric flows

Current numerical oceanic and atmospheric models suffer from a number of well-identified problems. These problems are mainly related to lack of horizontal and vertical resolution, thus requiring the parameterization of unresolved (subgrid scale) processes and control of discretization errors in order to fulfill criteria related to the particular underlying physics of rotating and strongly stratified flows. Oceanic and atmospheric coupled models are increasingly used in a wide range of applications from global to regional scales. Assessment of the reliability of those coupled models is an emerging topic as the spread among the solutions of existing models (e.g., for climate change predictions) has not been reduced with the new generation models when compared to the older ones.

Advanced methods for modeling 3D rotating and stratified flows The continuous increase of computational power and the resulting finer grid resolutions have triggered a recent regain of interest in numerical methods and their relation to physical processes. Going beyond present knowledge requires a better understanding of numerical dispersion/dissipation ranges and their connection to model fine scales. Removing the leading order truncation error of numerical schemes is thus an active topic of research and each mathematical tool has to adapt to the characteristics of three dimensional stratified and rotating flows. Studying the link between discretization errors and subgrid scale parameterizations is also arguably one of the main challenges.

Complexity of the geometry, boundary layers, strong stratification and lack of resolution are the main sources of discretization errors in the numerical simulation of geophysical flows. This emphasizes the importance of the definition of the computational grids (and coordinate systems) both in horizontal and vertical directions, and the necessity of truly multi resolution approaches. At the same time, the role of the small scale dynamics on large scale circulation has to be taken into account. Such parameterizations may be of deterministic as well as stochastic nature and both approaches are taken by the AIRSEA team. The design of numerical schemes consistent with the parameterizations is also arguably one of the main challenges for the coming years. This work is complementary and linked to that on parameters estimation described in 3.4 .

Ocean Atmosphere interactions and formulation of coupled models State-of-the-art climate models (CMs) are complex systems under continuous development. A fundamental aspect of climate modeling is the representation of air-sea interactions. This covers a large range of issues: parameterizations of atmospheric and oceanic boundary layers, estimation of air-sea fluxes, time-space numerical schemes, non conforming grids, coupling algorithms ...Many developments related to these different aspects were performed over the last 10-15 years, but were in general conducted independently of each other.

The aim of our work is to revisit and enrich several aspects of the representation of air-sea interactions in CMs, paying special attention to their overall consistency with appropriate mathematical tools. We intend to work consistently on the physics and numerics. Using the theoretical framework of global-in-time Schwarz methods, our aim is to analyze the mathematical formulation of the parameterizations in a coupling perspective. From this study, we expect improved predictability in coupled models (this aspect will be studied using techniques described in 3.4). Complementary work on space-time nonconformities and acceleration of convergence of Schwarz-like iterative methods (see 6.1.2) are also conducted.

3.3. Model reduction / multiscale algorithms

The high computational cost of the applications is a common and major concern to have in mind when deriving new methodological approaches. This cost increases dramatically with the use of sensitivity analysis or parameter estimation methods, and more generally with methods that require a potentially large number of model integrations.

A dimension reduction, using either stochastic or deterministic methods, is a way to reduce significantly the number of degrees of freedom, and therefore the calculation time, of a numerical model.

Model reduction Reduction methods can be deterministic (proper orthogonal decomposition, other reduced bases) or stochastic (polynomial chaos, Gaussian processes, kriging), and both fields of research are very active. Choosing one method over another strongly depends on the targeted application, which can be as varied as real-time computation, sensitivity analysis (see e.g., section 6.4) or optimisation for parameter estimation (see below).

Our goals are multiple, but they share a common need for certified error bounds on the output. Our team has a 4-year history of working on certified reduction methods and has a unique positioning at the interface between deterministic and stochastic approaches. Thus, it seems interesting to conduct a thorough comparison of the two alternatives in the context of sensitivity analysis. Efforts will also be directed toward the development of efficient greedy algorithms for the reduction, and the derivation of goal-oriented sharp error bounds for non linear models and/or non linear outputs of interest. This will be complementary to our work on the deterministic reduction of parametrized viscous Burgers and Shallow Water equations where the objective is to obtain sharp error bounds to provide confidence intervals for the estimation of sensitivity indices.

Reduced models for coupling applications Global and regional high-resolution oceanic models are either coupled to an atmospheric model or forced at the air-sea interface by fluxes computed empirically preventing proper physical feedback between the two media. Thanks to high-resolution observational studies, the existence of air-sea interactions at oceanic mesoscales (i.e., at $\mathcal{O}(1km)$ scales) have been unambiguously shown. Those interactions can be represented in coupled models only if the oceanic and atmospheric models are run on the same high-resolution computational grid, and are absent in a forced mode. Fully coupled models

at high-resolution are seldom used because of their prohibitive computational cost. The derivation of a reduced model as an alternative between a forced mode and the use of a full atmospheric model is an open problem.

Multiphysics coupling often requires iterative methods to obtain a mathematically correct numerical solution. To mitigate the cost of the iterations, we will investigate the possibility of using reduced-order models for the iterative process. We will consider different ways of deriving a reduced model: coarsening of the resolution, degradation of the physics and/or numerical schemes, or simplification of the governing equations. At a mathematical level, we will strive to study the well-posedness and the convergence properties when reduced models are used. Indeed, running an atmospheric model at the same resolution as the ocean model is generally too expensive to be manageable, even for moderate resolution applications. To account for important fine-scale interactions in the computation of the air-sea boundary condition, the objective is to derive a simplified boundary layer model that is able to represent important 3D turbulent features in the marine atmospheric boundary layer.

Reduced models for multiscale optimization The field of multigrid methods for optimisation has known a tremendous development over the past few decades. However, it has not been applied to oceanic and atmospheric problems apart from some crude (non-converging) approximations or applications to simplified and low dimensional models. This is mainly due to the high complexity of such models and to the difficulty in handling several grids at the same time. Moreover, due to complex boundaries and physical phenomena, the grid interactions and transfer operators are not trivial to define.

Multigrid solvers (or multigrid preconditioners) are efficient methods for the solution of variational data assimilation problems. We would like to take advantage of these methods to tackle the optimization problem in high dimensional space. High dimensional control space is obtained when dealing with parameter fields estimation, or with control of the full 4D (space time) trajectory. It is important since it enables us to take into account model errors. In that case, multigrid methods can be used to solve the large scales of the problem at a lower cost, this being potentially coupled with a scale decomposition of the variables themselves.

3.4. Dealing with uncertainties

There are many sources of uncertainties in numerical models. They are due to imperfect external forcing, poorly known parameters, missing physics and discretization errors. Studying these uncertainties and their impact on the simulations is a challenge, mostly because of the high dimensionality and non-linear nature of the systems. To deal with these uncertainties we work on three axes of research, which are linked: sensitivity analysis, parameter estimation and risk assessment. They are based on either stochastic or deterministic methods.

Sensitivity analysis Sensitivity analysis (SA), which links uncertainty in the model inputs to uncertainty in the model outputs, is a powerful tool for model design and validation. First, it can be a pre-stage for parameter estimation (see 3.4), allowing for the selection of the more significant parameters. Second, SA permits understanding and quantifying (possibly non-linear) interactions induced by the different processes defining e.g., realistic ocean atmosphere models. Finally SA allows for validation of models, checking that the estimated sensitivities are consistent with what is expected by the theory. On ocean, atmosphere and coupled systems, only first order deterministic SA are performed, neglecting the initialization process (data assimilation). AIRSEA members and collaborators proposed to use second order information to provide consistent sensitivity measures, but so far it has only been applied to simple academic systems. Metamodels are now commonly used, due to the cost induced by each evaluation of complex numerical models: mostly Gaussian processes, whose probabilistic framework allows for the development of specific adaptive designs, and polynomial chaos not only in the context of intrusive Galerkin approaches but also in a black-box approach. Until recently, global SA was based primarily on a set of engineering practices. New mathematical and methodological developments have led to the numerical computation of Sobol' indices, with confidence intervals assessing for both metamodel and estimation errors. Approaches have also been extended to the case of dependent entries, functional inputs and/or output and stochastic numerical codes. Other types of indices and generalizations of Sobol' indices have also been introduced.

Concerning the stochastic approach to SA we plan to work with parameters that show spatio-temporal dependencies and to continue toward more realistic applications where the input space is of huge dimension with highly correlated components. Sensitivity analysis for dependent inputs also introduces new challenges. In our applicative context, it would seem prudent to carefully learn the spatio-temporal dependences before running a global SA. In the deterministic framework we focus on second order approaches where the sought sensitivities are related to the optimality system rather than to the model; i.e., we consider the whole forecasting system (model plus initialization through data assimilation).
All these methods allow for computing sensitivities and more importantly a posteriori error statistics.

Parameter estimation Advanced parameter estimation methods are barely used in ocean, atmosphere and coupled systems, mostly due to a difficulty of deriving adequate response functions, a lack of knowledge of these methods in the ocean-atmosphere community, and also to the huge associated computing costs. In the presence of strong uncertainties on the model but also on parameter values, simulation and inference are closely associated. Filtering for data assimilation and Approximate Bayesian Computation (ABC) are two examples of such association.
Stochastic approach can be compared with the deterministic approach, which allows to determine the sensitivity of the flow to parameters and optimize their values relying on data assimilation. This approach is already shown to be capable of selecting a reduced space of the most influent parameters in the local parameter space and to adapt their values in view of correcting errors committed by the numerical approximation. This approach assumes the use of automatic differentiation of the source code with respect to the model parameters, and optimization of the obtained raw code.

AIRSEA assembles all the required expertise to tackle these difficulties. As mentioned previously, the choice of parameterization schemes and their tuning has a significant impact on the result of model simulations. Our research will focus on parameter estimation for parameterized Partial Differential Equations (PDEs) and also for parameterized Stochastic Differential Equations (SDEs). Deterministic approaches are based on optimal control methods and are local in the parameter space (i.e., the result depends on the starting point of the estimation) but thanks to adjoint methods they can cope with a large number of unknowns that can also vary in space and time. Multiscale optimization techniques as described in 6.3 will be one of the tools used. This in turn can be used either to propose a better (and smaller) parameter set or as a criterion for discriminating parameterization schemes. Statistical methods are global in the parameter state but may suffer from the curse of dimensionality. However, the notion of parameter can also be extended to functional parameters. We may consider as parameter a functional entity such as a boundary condition on time, or a probability density function in a stationary regime. For these purposes, non-parametric estimation will also be considered as an alternative.

Risk assessment Risk assessment in the multivariate setting suffers from a lack of consensus on the choice of indicators. Moreover, once the indicators are designed, it still remains to develop estimation procedures, efficient even for high risk levels. Recent developments for the assessment of financial risk have to be considered with caution as methods may differ pertaining to general financial decisions or environmental risk assessment. Modeling and quantifying uncertainties related to extreme events is of central interest in environmental sciences. In relation to our scientific targets, risk assessment is very important in several areas: hydrological extreme events, cyclone intensity, storm surges...Environmental risks most of the time involve several aspects which are often correlated. Moreover, even in the ideal case where the focus is on a single risk source, we have to face the temporal and spatial nature of environmental extreme events. The study of extremes within a spatio-temporal framework remains an emerging field where the development of adapted statistical methods could lead to major progress in terms of geophysical understanding and risk assessment thus coupling data and model information for risk assessment.

Based on the above considerations we aim to answer the following scientific questions: how to measure risk in a multivariate/spatial framework? How to estimate risk in a non stationary context? How to reduce dimension (see 3.3) for a better estimation of spatial risk?

Extreme events are rare, which means there is little data available to make inferences of risk measures. Risk assessment based on observation therefore relies on multivariate extreme value theory. Interacting particle systems for the analysis of rare events is commonly used in the community of computer experiments. An open question is the pertinence of such tools for the evaluation of environmental risk.
Most numerical models are unable to accurately reproduce extreme events. There is therefore a real need to develop efficient assimilation methods for the coupling of numerical models and extreme data.

3.5. High performance computing

Methods for sensitivity analysis, parameter estimation and risk assessment are extremely costly due to the necessary number of model evaluations. This number of simulations require considerable computational resources, depends on the complexity of the application, the number of input variables and desired quality of approximations. To this aim, the AIRSEA team is an intensive user of HPC computing platforms, particularly grid computing platforms. The associated grid deployment has to take into account the scheduling of a huge number of computational requests and the links with data-management between these requests, all of these as automatically as possible. In addition, there is an increasing need to propose efficient numerical algorithms specifically designed for new (or future) computing architectures and this is part of our scientific objectives. According to the computational cost of our applications, the evolution of high performance computing platforms has to be taken into account for several reasons. While our applications are able to exploit space parallelism to its full extent (oceanic and atmospheric models are traditionally based on a spatial domain decomposition method), the spatial discretization step size limits the efficiency of traditional parallel methods. Thus the inherent parallelism is modest, particularly for the case of relative coarse resolution but with very long integration time (e.g., climate modeling). Paths toward new programming paradigms are thus needed. As a step in that direction, we plan to focus our research on parallel in time methods.

New numerical algorithms for high performance computing Parallel in time methods can be classified into three main groups. In the first group, we find methods using parallelism across the method, such as parallel integrators for ordinary differential equations. The second group considers parallelism across the problem. Falling into this category are methods such as waveform relaxation where the space-time system is decomposed into a set of subsystems which can then be solved independently using some form of relaxation techniques or multigrid reduction in time. The third group of methods focuses on parallelism across the steps. One of the best known algorithms in this family is parareal. Other methods combining the strengths of those listed above (e.g., PFASST) are currently under investigation in the community.

Parallel in time methods are iterative methods that may require a large number of iteration before convergence. Our first focus will be on the convergence analysis of parallel in time (Parareal / Schwarz) methods for the equation systems of oceanic and atmospheric models. Our second objective will be on the construction of fast (approximate) integrators for these systems. This part is naturally linked to the model reduction methods of section (6.3.1). Fast approximate integrators are required both in the Schwarz algorithm (where a first guess of the boundary conditions is required) and in the Parareal algorithm (where the fast integrator is used to connect the different time windows). Our main application of these methods will be on climate (i.e., very long time) simulations. Our second application of parallel in time methods will be in the context of optimization methods. In fact, one of the major drawbacks of the optimal control techniques used in 3.4 is a lack of intrinsic parallelism in comparison with ensemble methods. Here, parallel in time methods also offer ways to better efficiency. The mathematical key point is centered on how to efficiently couple two iterative methods (i.e., parallel in time and optimization methods).

ANGE Project-Team

3. Research Program

3.1. Overview

The research activities carried out within the ANGE team strongly couple the development of methodological tools with applications to real–life problems and the transfer of numerical codes. The main purpose is to obtain new models adapted to the physical phenomena at stake, identify the main properties that reflect the physical meaning of the models (uniqueness, conservativity, entropy dissipation, ...), propose effective numerical methods to approximate their solution in complex configurations (multi-dimensional, unstructured meshes, well-balanced, ...) and to assess the results with data in the purpose of potentially correcting the models.

The difficulties arising in gravity driven flow studies are threefold.

- Models and equations encountered in fluid mechanics (typically the free surface Navier-Stokes equations) are complex to analyze and solve.
- The underlying phenomena often take place over large domains with very heterogeneous length scales (size of the domain, mean depth, wave length, ...) and distinct time scales, *e.g.* coastal erosion, propagation of a tsunami, ...
- These problems are multi-physics with strong couplings and nonlinearities.

3.2. Modelling and analysis

Hazardous flows are complex physical phenomena that can hardly be represented by shallow water type systems of partial differential equations (PDEs). In this domain, the research program is devoted to the derivation and analysis of reduced complexity models compared to the Navier-Stokes equations, but relaxing the shallow water assumptions. The main purpose is then to obtain models well-adapted to the physical phenomena at stake.

Even if the resulting models do not strictly belong to the family of hyperbolic systems, they exhibit hyperbolic features: the analysis and discretisation techniques we intend to develop have connections with those used for hyperbolic conservation laws. It is worth noticing that the need for robust and efficient numerical procedures is reinforced by the smallness of dissipative effects in geophysical models which therefore generate singular solutions and instabilities.

On the one hand, the derivation of the Saint-Venant system from the Navier-Stokes equations is based on two approximations (the so-called shallow water assumptions), namely

- the horisontal fluid velocity is well approximated by its mean value along the vertical direction,
- the pressure is hydrostatic or equivalently the vertical acceleration of the fluid can be neglected compared to the gravitational effects.

As a consequence the objective is to get rid of these two assumptions, one after the other, in order to obtain models accurately approximating the incompressible Euler or Navier-Stokes equations.

On the other hand, many applications require the coupling with non-hydrodynamic equations, as in the case of micro-algae production or erosion processes. These new equations comprise non-hyperbolic features and a special analysis is needed.

3.2.1. Multilayer approach

As for the first shallow water assumption, *multi-layer* systems were proposed to describe the flow as a superposition of Saint-Venant type systems [26], [29], [30]. Even if this approach has provided interesting results, layers are considered separate and non-miscible fluids, which implies strong limitations. That is why we proposed a slightly different approach [27], [28] based on a Galerkin type decomposition along the vertical axis of all variables and leading, both for the model and its discretisation, to more accurate results.

A kinetic representation of our multilayer model allows to derive robust numerical schemes endowed with crucial properties such as: consistency, conservativity, positivity, preservation of equilibria, ... It is one of the major achievements of the team but it needs to be analyzed and extended in several directions namely:

- The convergence of the multilayer system towards the hydrostatic Euler system as the number of layers goes to infinity is a critical point. It is not fully satisfactory to have only formal estimates of the convergence and sharp estimates would provide an optimal number of layers.
- The introduction of several source terms due for instance to the Coriolis force or extra terms from changes of coordinates seems necessary. Their inclusion should lead to substantial modifications of the numerical scheme.
- Its hyperbolicity has not yet been proven and conversely the possible loss of hyperbolicity cannot be characterised. Similarly, the hyperbolic feature is essential in the propagation and generation of waves.

3.2.2. Non-hydrostatic models

The hydrostatic assumption consists in neglecting the vertical acceleration of the fluid. It is considered valid for a large class of geophysical flows but is restrictive in various situations where the dispersive effects (like wave propagation) cannot be neglected. For instance, when a wave reaches the coast, bathymetry variations give a vertical acceleration to the fluid that strongly modifies the wave characteristics and especially its height.

Processing an asymptotic expansion (w.r.t. the aspect ratio for shallow water flows) into the Navier-Stokes equations, we obtain at the leading order the Saint-Venant system. Going one step further leads to a vertically averaged version of the Euler/Navier-Stokes equations involving some non-hydrostatic terms. This model has several advantages:

- it admits an energy balance law (that is not the case for most dispersive models available in the literature),
- it reduces to the Saint-Venant system when the non-hydrostatic pressure term vanishes,
- it consists in a set of conservation laws with source terms,
- it does not contain high order derivatives.

3.2.3. Multi-physics modelling

The coupling of hydrodynamic equations with other equations in order to model interactions between complex systems represents an important part of the team research. More precisely, three multi-physics systems are investigated. More details about the industrial impact of these studies are presented in the following section.

- To estimate the risk for infrastructures in coastal zones or close to a river, the resolution of the shallow water equations with moving bathymetry is necessary. The first step consisted in the study of an additional equation largely used in engineering science: The Exner equation. The analysis enabled to exhibit drawbacks of the coupled model such as the lack of energy conservation or the strong variations of the solution from small perturbations. A new formulation is proposed to avoid these drawbacks. The new model consists in a coupling between conservation laws and an elliptic equation, like the Euler/Poisson system, suggesting to use well-known strategies for the analysis and the numerical resolution. In addition, the new formulation is derived from classical complex rheology models and allowed physical phenomena like threshold laws.
- Interaction between flows and floating structures is the challenge at the scale of the shallow water equations. This study requires a better understanding of the energy exchanges between the flow and the structure. The mathematical model of floating structures is very hard to solve numerically due to the non-penetration condition at the interface between the flow and the structure. It leads to infinite potential wave speeds that could not be solved with classical free surface numerical schemes. A relaxation model was derived to overcome this difficulty. It represents the interaction with the floating structure with a free surface model-type.

- If the interactions between hydrodynamics and biology phenomena are known through laboratory experiments, it is more difficult to predict the evolution, especially for the biological quantities, in a real and heterogeneous system. The objective is to model and reproduce the hydrodynamics modifications due to forcing term variations (in time and space). We are typically interested in phenomena such as eutrophication, development of harmful bacteria (cyanobacteria) and upwelling phenomena.

3.2.4. Data assimilation and inverse modelling

In environmental applications, the most accurate numerical models remain subject to uncertainties that originate from their parameters and shortcomings in their physical formulations. It is often desirable to quantify the resulting uncertainties in a model forecast. The propagation of the uncertainties may require the generation of ensembles of simulations that ideally sample from the probability density function of the forecast variables. Classical approaches rely on multiple models and on Monte Carlo simulations. The applied perturbations need to be calibrated for the ensemble of simulations to properly sample the uncertainties. Calibrations involve ensemble scores that compare the consistency between the ensemble simulations and the observational data. The computational requirements are so high that designing fast surrogate models or metamodels is often required.

In order to reduce the uncertainties, the fixed or mobile observations of various origins and accuracies can be merged with the simulation results. The uncertainties in the observations and their representativeness also need to be quantified in the process. The assimilation strategy can be formulated in terms of state estimation or parameter estimation (also called inverse modelling). Different algorithms are employed for static and dynamic models, for analyses and forecasts. A challenging question lies in the optimization of the observational network for the assimilation to be the most efficient at a given observational cost.

3.3. Numerical analysis

3.3.1. Non-hydrostatic scheme

The main challenge in the study of the non-hydrostatic model is to design a robust and efficient numerical scheme endowed with properties such as: positivity, wet/dry interfaces treatment, consistency. It must be noticed that even if the non-hydrostatic model looks like an extension of the Saint-Venant system, most of the known techniques used in the hydrostatic case are not efficient as we recover strong difficulties encountered in incompressible fluid mechanics due to the extra pressure term. These difficulties are reinforced by the absence of viscous/dissipative terms.

3.3.2. Space decomposition and adaptive scheme

In the quest for a better balance between accuracy and efficiency, a strategy consists in the adaptation of models. Indeed, the systems of partial differential equations we consider result from a hierarchy of simplifying assumptions. However, some of these hypotheses may turn out to be irrelevant locally. The adaptation of models thus consists in determining areas where a simplified model (*e.g.* shallow water type) is valid and where it is not. In the latter case, we may go back to the "parent" model (*e.g.* Euler) in the corresponding area. This implies to know how to handle the coupling between the aforementioned models from both theoretical and numerical points of view. In particular, the numerical treatment of transmission conditions is a key point. It requires the estimation of characteristic values (Riemann invariant) which have to be determined according to the regime (torrential or fluvial).

3.3.3. Asymptotic-Preserving scheme for source terms

Hydrodynamic models comprise advection and sources terms. The conservation of the balance between source terms, typically viscosity and friction, has a significant impact since the overall flow is generally a perturbation around an equilibrium. The design of numerical schemes able to preserve such balances is a challenge from both theoretical and industrial points of view. The concept of Asymptotic-Preserving (AP) methods is of great interest in order to overcome these issues.

Another difficulty occurs when a term, typically related to the pressure, becomes very large compared to the order of magnitude of the velocity. At this regime, namely the so-called *low Froude* (shallow water) or *low Mach* (Euler) regimes, the difference between the speed of the gravity waves and the physical velocity makes classical numerical schemes inefficient: firstly because of the error of truncation which is inversely proportional to the small parameters, secondly because of the time step governed by the largest speed of the gravity wave. AP methods made a breakthrough in the numerical resolution of asymptotic perturbations of partial-differential equations concerning the first point. The second one can be fixed using partially implicit scheme.

3.3.4. Multi-physics models

Coupling problems also arise within the fluid when it contains pollutants, density variations or biological species. For most situations, the interactions are small enough to use a splitting strategy and the classical numerical scheme for each sub-model, whether it be hydrodynamic or non-hydrodynamic.

The sediment transport raises interesting issues from a numerical aspect. This is an example of coupling between the flow and another phenomenon, namely the deformation of the bottom of the basin that can be carried out either by bed load where the sediment has its own velocity or suspended load in which the particles are mostly driven by the flow. This phenomenon involves different time scales and nonlinear retroactions; hence the need for accurate mechanical models and very robust numerical methods. In collaboration with industrial partners (EDF–LNHE), the team already works on the improvement of numerical methods for existing (mostly empirical) models but our aim is also to propose new (quite) simple models that contain important features and satisfy some basic mechanical requirements. The extension of our 3D models to the transport of weighted particles can also be here of great interest.

3.3.5. Optimisation

Numerical simulations are a very useful tool for the design of new processes, for instance in renewable energy or water decontamination. The optimisation of the process according to a well-defined objective such as the production of energy or the evaluation of a pollutant concentration is the logical upcoming challenge in order to propose competitive solutions in industrial context. First of all, the set of parameters that have a significant impact on the result and on which we can act in practice is identified. Then the optimal parameters can be obtained using the numerical codes produced by the team to estimate the performance for a given set of parameters with an additional loop such as gradient descent or Monte Carlo method. The optimisation is used in practice to determine the best profile for turbine pales, the best location for water turbine implantation, in particular for a farm.

ARAMIS Project-Team

3. Research Program

3.1. From geometrical data to multimodal imaging

Brain diseases are associated to alterations of brain structure that can be studied in vivo using anatomical and diffusion MRI. The anatomy of a given subject can be represented by sets of anatomical surfaces (cortical and subcortical surfaces) and curves (white matter tracks) that can be extracted from anatomical and diffusion MRI respectively. We aim to develop approaches that can characterize the variability of brain anatomy within populations of subjects. To that purpose, we propose methods to estimate population atlases that provide an average model of a population of subjects together with a statistical model of their variability. Finally, we aim to introduce representations that can integrate geometrical information (anatomical surfaces, white matter tracts) together with functional (PET, ASL, EEG/MEG) and microstructural information.

3.2. Models of brain networks

Functional imaging techniques (EEG, MEG and fMRI) allow characterizing the statistical interactions between the activities of different brain areas, i.e. functional connectivity. Functional integration of spatially distributed brain regions is a well-known mechanism underlying various cognitive tasks, and is disrupted in brain disorders. Our team develops a framework for the characterization of brain connectivity patterns, based on connectivity descriptors from the theory of complex networks. More specifically, we propose analytical tools to infer brain networks, chacterize their structure and integrate multiple networks (for instance from multiple frequency bands or multiple modalities). The genericity of this approach allows us to apply it to various types of data including functional and structural neuroimaging, as well as genomic data.

3.3. Spatiotemporal modeling from longitudinal data

Longitudinal data sets are collected to capture variable temporal phenomena, which may be due to ageing or disease progression for instance. They consist in the observation of several individuals, each of them being observed at multiple points in time. The statistical exploitation of such data sets is notably difficult since data of each individual follow a different trajectory of changes and at its own pace. This difficulty is further increased if observations take the form of structured data like images or measurements distributed at the nodes of a mesh, and if the measurements themselves are normalized data or positive definite matrices for which usual linear operations are not defined. We aim to develop a theoretical and algorithmic framework for learning typical trajectories from longitudinal data sets. This framework is built on tools from Riemannian geometry to describe trajectories of changes for any kind of data and their variability within a group both in terms of the direction of the trajectories and pace.

3.4. Decision support systems

We then aim to develop tools to assist clinical decisions such as diagnosis, prognosis or inclusion in therapeutic trials. To that purpose, we leverage the tools developed by the team, such as multimodal representations, network indices and spatio-temporal models which are combined with advanced classification and regression approaches. We also dedicate strong efforts to rigorous, transparent and reproducible validation of the decision support systems on large clinical datasets.

3.5. Clinical research studies

Finally, we aim to apply advanced computational and statistical tools to clinical research studies. These studies are often performed in collaboration with other researchers of the ICM, clinicians of the Pitié -Salpêtrière hospital or external partners. Notably, our team is very often involved ”ex-ante” in clinical research studies. As co-investigators of such studies, we contribute to the definition of objectives, study design and definition of protocols. This is instrumental to perform clinically relevant methodological development and to maximize their medical impact. A large part of these clinical studies were in the field of dementia (Alzheimer's disease, fronto-temporal dementia). Recently, we expanded our scope to other neurodegenerative diseases (Parkinson's disease, multiple sclerosis).

ATHENA Project-Team

3. Research Program

3.1. Computational diffusion MRI

Diffusion MRI (dMRI) provides a non-invasive way of estimating in-vivo CNS fiber structures using the average random thermal movement (diffusion) of water molecules as a probe. It's a relatively recent field of research with a history of roughly three decades. It was introduced in the mid 80's by Le Bihan et al [63], Merboldt et al [68] and Taylor et al [78]. As of today, it is the unique non-invasive technique capable of describing the neural connectivity in vivo by quantifying the anisotropic diffusion of water molecules in biological tissues.

3.1.1. Diffusion Tensor Imaging & High Angular Resolution Diffusion Imaging

In dMRI, the acquisition and reconstruction of the diffusion signal allows for the reconstruction of the water molecules displacement probability, known as the Ensemble Average Propagator (EAP) [77], [47]. Historically, the first model in dMRI is the 2nd order diffusion tensor (DTI) [45], [44] which assumes the EAP to be Gaussian centered at the origin. DTI (Diffusion Tensor Imaging) has now proved to be extremely useful to study the normal and pathological human brain [64], [55]. It has led to many applications in clinical diagnosis of neurological diseases and disorder, neurosciences applications in assessing connectivity of different brain regions, and more recently, therapeutic applications, primarily in neurosurgical planning. An important and very successful application of diffusion MRI has been brain ischemia, following the discovery that water diffusion drops immediately after the onset of an ischemic event, when brain cells undergo swelling through cytotoxic edema.

The increasing clinical importance of diffusion imaging has driven our interest to develop new processing tools for Diffusion Tensor MRI. Because of the complexity of the data, this imaging modality raises a large amount of mathematical and computational challenges. We have therefore developed original and efficient algorithms relying on Riemannian geometry, differential geometry, partial differential equations and front propagation techniques to correctly and efficiently estimate, regularize, segment and process Diffusion Tensor MRI (DT-MRI) (see [66] and [65]).

In DTI, the Gaussian assumption over-simplifies the diffusion of water molecules. While it is adequate for voxels in which there is only a single fiber orientation (or none), it breaks for voxels in which there are more complex internal structures and limitates the ability of the DTI to describe complex, singular and intricate fiber configurations (U-shape, kissing or crossing fibers). To overcome this limitation, so-called Diffusion Spectrum Imaging (DSI) [81] and High Angular Resolution Diffusion Imaging (HARDI) methods such as Q-ball imaging [79] and other multi-tensors and compartment models [74], [76], [38], [37], [71] were developed to resolve the orientationnality of more complicated fiber bundle configurations.

Q-Ball imaging (QBI) has been proven very successful in resolving multiple intravoxel fiber orientations in MR images, thanks to its ability to reconstruct the Orientation Distribution Function (ODF, the probability of diffusion in a given direction). These tools play a central role in our work related to the development of a robust and linear spherical harmonic estimation of the HARDI signal and to our development of a regularized, fast and robust analytical QBI solution that outperforms the state-of-the-art ODF numerical technique developed by Tuch [79]. Those contributions are fundamental and have already started to impact on the Diffusion MRI, HARDI and Q-Ball Imaging community [54]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [51], [3] and [52], [4]).

3.1.2. Beyond DTI with high order tensors

High Order Tensors (HOT) models to estimate the diffusion function while overcoming the shortcomings of the 2nd order tensor model have also been proposed such as the Generalized Diffusion Tensor Imaging (G-DTI) model developed by Ozarslan et al [85], [86] or 4th order Tensor Model [43]. For more details, we refer the reader to our articles in [57], [74] where we review HOT models and to our articles in [65], co-authored with some of our close collaborators, where we review recent mathematical models and computational methods for the processing of Diffusion Magnetic Resonance Images, including state-of-the-art reconstruction of diffusion models, cerebral white matter connectivity analysis, and segmentation techniques. We also worked on Diffusion Kurtosis Imaging (DKI), of great interest for the company OLEA MEDICAL(https://www.olea-medical.com/en). Indeed, DKI is fastly gaining popularity in the domain for characterizing the diffusion propagator or EAP by its deviation from Gaussianity. Hence it is an important clinical tool for characterizing the white-matter's integrity with biomarkers derived from the 3D 4th order kurtosis tensor (KT) [60].

All these powerful techniques are of utmost importance to acquire a better understanding of the CNS mechanisms and have helped to efficiently tackle and solve a number of important and challenging problems [37], [38]. They have also opened up a landscape of extremely exciting research fields for medicine and neuroscience. Hence, due to the complexity of the CNS data and as the magnetic field strength of scanners increases, as the strength and speed of gradients increase and as new acquisition techniques appear [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [59], [74].

3.1.3. Improving dMRI acquisitions

One of the most important challenges in diffusion imaging is to improve acquisition schemes and analyse approaches to optimally acquire and accurately represent diffusion profiles in a clinically feasible scanning time. Indeed, a very important and open problem in Diffusion MRI is related to the fact that HARDI scans generally require many times more diffusion gradient than traditional diffusion MRI scan times. This comes at the price of longer scans, which can be problematic for children and people with certain diseases. Patients are usually unable to tolerate long scans and excessive motion of the patient during the acquisition process can force a scan to be aborted or produce useless diffusion MRI images. We have developed novel methods for the acquisition and the processing of diffusion magnetic resonance images, to efficiently provide, with just few measurements, new insights into the structure and anatomy of the brain white matter in vivo.

First, we contributed developing real-time reconstruction algorithm based on the Kalman filter [50]. Then, we started to explore the utility of Compressive Sensing methods to enable faster acquisition of dMRI data by reducing the number of measurements, while maintaining a high quality for the results. Compressed Sensing (CS) is a relatively recent technique which has been proved to accurately reconstruct sparse signals from undersampled measurements acquired below the Shannon-Nyquist rate [69].

We have contributed to the reconstruction of the diffusion signal and its important features as the orientation distribution function and the ensemble average propagator, with a special focus on clinical setting in particular for single and multiple Q-shell experiments. Compressive sensing as well as the parametric reconstruction of the diffusion signal in a continuous basis of functions such as the Spherical Polar Fourier basis, have been proved through our contributions to be very useful for deriving simple and analytical closed formulae for many important dMRI features, which can be estimated via a reduced number of measurements [69], [48], [49].

We have also contributed to design optimal acquisition schemes for single and multiple Q-shell experiments. In particular, the method proposed in [2] helps generate sampling schemes with optimal angular coverage for multi-shell acquisitions. The cost function we proposed is an extension of the electrostatic repulsion to multi-shell and can be used to create acquisition schemes with incremental angular distribution, compatible with prematurely stopped scans. Compared to more commonly used radial sampling, our method improves the angular resolution, as well as fiber crossing discrimination. The optimal sampling schemes, freely available for download[0], have been selected for use in the HCP (Human Connectome Project)[0].

[0] http://www.emmanuelcaruyer.com/
[0] http://humanconnectome.org/documentation/Q1/imaging-protocols.html

We think that such kind of contributions open new perspectives for dMRI applications including, for example, tractography where the improved characterization of the fiber orientations is likely to greatly and quickly help tracking through regions with and/or without crossing fibers [58].

3.1.4. dMRI modelling, tissue microstructures features recovery & applications

The dMRI signal is highly complex, hence, the mathematical tools required for processing it have to be commensurate in their complexity. Overall, these last twenty years have seen an explosion of intensive scientific research which has vastly improved and literally changed the face of dMRI. In terms of dMRI models, two trends are clearly visible today: the parametric approaches which attempt to build models of the tissue to explain the signal based on model-parameters such as CHARMED [39], AxCaliber [40] and NODDI [82] to cite but a few, and the non-parametric approaches, which attempt to describe the signal in useful but generic functional bases such as the Spherical Polar Fourier (SPF) basis [42], [41], the Solid Harmonic (SoH) basis [53], the Simple Harmonic Oscillator based Reconstruction and Estimation (SHORE) basis [83] and more recent Mean Apparent Propagator or MAP-MRI basis [84].

We propose to investigate the feasibility of using our new models and methods to measure extremely important biological tissue microstructure quantities such as axonal radius and density in white matter. These parameters could indeed provide new insight to better understand the brain's architecture and more importantly could also provide new imaging bio-markers to characterize certain neurodegenerative diseases. This challenging scientific problem, when solved, will lead to direct measurements of important microstructural features that will be integrated in our analysis to provide much greater insight into disease mechanisms, recovery and development. These new microstructural parameters will open the road to go far beyond the limitations of the more simple bio-markers derived from DTI that are clinically used to this date – such as MD (Mean Diffusivity) and FA (Fractional Anisotropy) which are known to be extremely sensitive to confounding factors such as partial volume and axonal dispersion, non-specific and not able to capture any subtle effects that might be early indicators of diseases [5].

3.1.5. Towards microstructural based tractography

In order to go far beyond traditional fiber-tracking techniques, we believe that first order information, i.e. fiber orientations, has to be superseeded by second and third order information, such as microstructure details, to improve tractography. However, many of these higher order information methods are relatively new or unexplored and tractography algorithms based on these high order based methods have to be conceived and designed. In this aim, we propose to work with multiple-shells to reconstruct the Ensemble Average Propagator (EAP), which represents the whole 3D diffusion process and use the possibility it offers to deduce valuable insights on the microstructural properties of the white matter. Indeed, from a reconstructed EAP one can compute the angular features of the diffusion in an diffusion Orientation Distribution Function (ODF), providing insight in axon orientation, calculate properties of the entire diffusion in a voxel such as the Mean Squared Diffusivity (MSD) and Return-To-Origin Probability (RTOP), or come forth with bio-markers detailing diffusion along a particular white matter bundle direction such as the Return-to-Axis or Return-to-Plane Probability (RTAP or RTPP). This opens the way to a ground-breaking computational and unified framework for tractography based on EAP and microstructure features [6]. Using additional a priori anatomical and/or functional information, we could also constrain the tractography algorithm to start and terminate the streamlines only at valid processing areas of the brain.

This development of a computational and unified framework for tractography, based on EAP, microstructure and a priori anatomical and/or functional features, will open new perspectives in tractography, paving the way to a new generation of realistic and biologically plausible algorithms able to deal with intricate configurations of white matter fibers and to provide an exquisite and intrinsic brain connectivity quantification.

3.1.6. Going beyond the state-of-the-art dMRI

Overall, these last twenty years have seen an explosion of intensive scientific research which has vastly improved and literally changed the face of dMRI.

However, although great improvements have been made, major improvements are still required primarily to optimally acquire dMRI data, better understand the biophysics of the signal formation, recover high order invariant and intrinsic microstructure features, identify bio-physically important bio-markers and improve tractography.

Therefore, there is still considerable room for improvement when it comes to the concepts and tools able to efficiently acquire, process and analyze the complex structure of dMRI data. Develop ground-breaking dMRI tools and models for brain connectomics is one of the major objective we would like to achieve in order to take dMRI from the benchside to the bedside and lead to a decisive advance and breakthrough in this field.

3.2. MEG and EEG

Electroencephalography (EEG) and Magnetoencephalography (MEG) are two non-invasive techniques for measuring (part of) the electrical activity of the brain. While EEG is an old technique (Hans Berger, a German neuropsychiatrist, measured the first human EEG in 1929), MEG is a rather new one: the first measurements of the magnetic field generated by the electrophysiological activity of the brain were made in 1968 at MIT by D. Cohen. Nowadays, EEG is relatively inexpensive and is routinely used to detect and qualify neural activities (epilepsy detection and characterisation, neural disorder qualification, BCI, ...). MEG is, comparatively, much more expensive as SQUIDS (Superconducting QUantum Interference Device) only operate under very challenging conditions (at liquid helium temperature) and as a specially shielded room must be used to separate the signal of interest from the ambient noise. However, as it reveals a complementary vision to that of EEG and as it is less sensitive to the head structure, it also bears great hopes and an increasing number of MEG machines are being installed throughout the world. Inria and ODYSSÉE/ATHENA have participated in the acquisition of one such machine installed in the hospital "La Timone" in Marseille.

MEG and EEG can be measured simultaneously (M/EEG) and reveal complementary properties of the electrical fields. The two techniques have temporal resolutions of about the millisecond, which is the typical granularity of the measurable electrical phenomena that arise within the brain. This high temporal resolution makes MEG and EEG attractive for the functional study of the brain. The spatial resolution, on the contrary, is somewhat poor as only a few hundred data points can be acquired simultaneously (about 300-400 for MEG and up to 256 for EEG). MEG and EEG are somewhat complementary with fMRI (Functional MRI) and SPECT (Single-Photon Emission Computed Tomography) in that those provide a very good spatial resolution but a rather poor temporal resolution (of the order of a second for fMRI and a minute for SPECT). Also, contrarily to fMRI, which “only” measures an haemodynamic response linked to the metabolic demand, MEG and EEG measure a direct consequence of the electrical activity of the brain: it is acknowledged that the signals measured by MEG and EEG correspond to the variations of the post-synaptic potentials of the pyramidal cells in the cortex. Pyramidal neurons compose approximately 80% of the neurons of the cortex, and it requires at least about 50,000 active such neurons to generate some measurable signal.

While the few hundred temporal curves obtained using M/EEG have a clear clinical interest, they only provide partial information on the localisation of the sources of the activity (as the measurements are made on or outside of the head). Thus the practical use of M/EEG data raises various problems that are at the core of the ATHENA research in this topic:

- First, as acquisition is continuous and is run at a rate up to 1kHz, the amount of data generated by each experiment is huge. Data selection and reduction (finding relevant time blocks or frequency bands) and pre-processing (removing artifacts, enhancing the signal to noise ratio, ...) are largely done manually at present. Making a better and more systematic use of the measurements is an important step to optimally exploit the M/EEG data [1].
- With a proper model of the head and of the sources of brain electromagnetic activity, it is possible to simulate the electrical propagation and reconstruct sources that can explain the measured signal. Proposing better models [62], [7] and means to calibrate them [80] so as to have better reconstructions are other important aims of our work.

- Finally, we wish to exploit the temporal resolution of M/EEG and to apply the various methods we have developed to better understand some aspects of the brain functioning, and/or to extract more subtle information out of the measurements. This is of interest not only as a cognitive goal, but it also serves the purpose of validating our algorithms and can lead to the use of such methods in the field of Brain Computer Interfaces. To be able to conduct such kind of experiments, an EEG lab has been set up at ATHENA.

3.3. Combined M/EEG and dMRI

dMRI provides a global and systematic view of the long-range structural connectivity within the whole brain. In particular, it allows the recovery of the fiber structure of the white matter which can be considered as the wiring connections between distant cortical areas. These white matter based tractograms are analyzed e.g. to explore the differences in structural connectivity between pathological and normal populations. Moreover, as a by-product, the tractograms can be processed to reveal the nodes of the brain networks, i.e. by segregating together gray matter that share similar connections to the rest of the white matter. But dMRI does not provide information on:

- the cortico-cortical pathways (not passing through white matter) and to some extent, on the short-range connections in the white matter,
- the actual use of connections over time during a given brain activity.

On the opposite, M/EEG measures brain activation over time and provides, after source reconstruction (solving the so-called inverse problem of source reconstruction), time courses of the activity of the cortical areas. Unfortunately, deep brain structures have very little contribution to M/EEG measurements and are thus difficult to analyze. Consequently, M/EEG reveals information about the nodes of the network, but in a more blurry (because of the inverse problem) and fragmented view than dMRI (since it can only reveal brain areas measurable in M/EEG whose activity varies during the experimental protocol). Given its very high temporal resolution, the signal of reconstructed sources can be processed to reveal the functional connectivity between the nodes [75].

While dMRI and M/EEG have been the object of considerable research separately, there have been very few studies on combining the information they provide. Some existing studies deal with the localization of abnormal MEG signals, particularly in the case of epilepsy, and on studying the white matter fibers near the detected abnormal source [67], [70], but to our knowledge there are very few studies merging data coming both from M/EEG and dMRI at the analysis level [72], [56], [46], [73].

Combining the structural and functional information provided by dMRI and M/EEG is a difficult problem as the spatial and temporal resolutions of the two types of measures are extremely different. Still, combining the measurements obtained by these two types of techniques has the great potential of providing a detailed view both in space and time of the functioning brain at a macroscopic level. Consequently, it is a timely and extremely important objective to develop innovative computational tools and models that advance the dMRI and M/EEG state-of-the-art and combine these imaging modalities to build a comprehensive dynamical structural-functional brain connectivity network to be exploited in brain connectivities diseases.

The COBCOM ERC project aims to develop a joint dynamical structural-functional brain connectivity network built on advanced and integrated dMRI and M/EEG ground-breaking methods. To this end, COBCOM will provide new generation of computational dMRI and M/EEG models and methods for identifying and characterizing the connectivities on which the joint network is built. Capitalizing on the strengths of dMRI & M/EEG and building on the bio-physical and mathematical foundations of our models, COBCOM will contribute to create a joint and solid network which will be exploited to identify and characterize white matter abnormalities in some high-impact brain diseases such as Multiple Sclerosis (MS), Epilepsy and mild Traumatic Brain Injury (mTBI).

BEAGLE Project-Team

3. Research Program

3.1. Introduction

As stated above, the research topics of the BEAGLE Team are centered on the modelization and simulation of cellular processes. More specifically, we focus on two specific processes that govern cell dynamics and behavior: Biophysics and Evolution. We are strongly engaged into the integration of these level of biological understanding.

3.2. Research axis 1: Computational cellular biochemistry

Biochemical kinetics developed as an extension of chemical kinetics in the early 20th century and inherited the main hypotheses underlying Van't Hoff's law of mass action : a perfectly-stirred homogeneous medium with deterministic kinetics. This classical view is however challenged by recent experimental results regarding both the movement and the metabolic fate of biomolecules. First, it is now known that the diffusive motion of many proteins in cellular media exhibits deviations from the ideal case of Brownian motion, in the form of position-dependent diffusion or anomalous diffusion, a hallmark of poorly mixing media. Second, several lines of evidence indicate that the metabolic fate of molecules in the organism not only depends on their chemical nature, but also on their spatial organisation – for example, the fate of dietary lipids depends on whether they are organized into many small or a few large droplets (see e.g. [28]). In this modern-day framework, cellular media appear as heterogeneous collections of contiguous spatial domains with different characteristics, thus providing spatial organization of the reactants. Moreover, the number of implicated reactants is often small enough that stochasticity cannot be ignored. To improve our understanding of intracellular biochemistry, we study spatiotemporal biochemical kinetics using computer simulations (particle-based spatially explicit stochastic simulations) and mathematical models (age-structured PDEs).

3.3. Research axis 2: Models for Molecular Evolution

We study the processes of genome evolution, with a focus on large-scale genomic events (rearrangements, duplications, transfers). We are interested in deciphering general laws which explain the organization of the genomes we observe today, as well as using the knowledge of these processes to reconstruct some aspects of the history of life. To do so, we construct mathematical models and apply them either in a "forward" way, *i.e.* observing the course of evolution from known ancestors and parameters, by simulation (*in silico experimental evolution*) or mathematical analysis (*theoretical biology*), or in a "backward" way, *i.e.* reconstructing ancestral states and parameters from known extant states (*phylogeny, comparative genomics*). Moreover we often mix the two approaches either by validating backwards reconstruction methods on forward simulations, or by using the forward method to test evolutionary hypotheses on biological data.

3.4. Research axis 3: Computational systems biology of neurons and astrocytes

Brain cells are rarely considered by computational systems biologists, though they are especially well suited for the field: their major signaling pathways are well characterized, the cellular properties they support are well identified (e.g. synaptic plasticity) and eventually give rise to well known functions at the organ scale (learning, memory). Moreover, electro-physiology measurements provide us with an experimental monitoring of signaling at the single cell level (sometimes at the sub-cellular scale) with unrivaled temporal resolution (milliseconds) over durations up to an hour. In this research axis, we develop modeling approaches for systems biology of both neuronal cells and glial cells, in particular astrocytes. We are mostly interested in understanding how the pathways implicated in the signaling between neurons, astrocytes and neurons-astrocytes interactions implement and regulate synaptic plasticity.

3.5. Research axis 4: Evolutionary Systems Biology

This axis, consisting in integrating the two main biological levels we study, is a long-standing and long-term objective in the team. These last years we did not make significant advances in this direction and we even removed this objective from last year's report. However the evolution of the team staff and projects allows us to give it back its central place. We now have the forces and ideas to progress. We have several short and middle term projects to integrate biochemical data and evolution. In particular we are analysing with an evolutionary perspective the 3D conformation of chromosomes, the regulatory landscape of genomes, the chromatin-associated proteins.

BIGS Project-Team

3. Research Program

3.1. Introduction

We give here the main lines of our research that belongs to the domains of probability and statistics. For clarity, we made the choice to structure them in four items. Although this choice was not arbitrary, the outlines between these items are sometimes fuzzy because each of them deals with modeling and inference and they are all interconnected.

3.2. Stochastic modeling

Our aim is to propose relevant stochastic frameworks for the modeling and the understanding of biological systems. The stochastic processes are particularly suitable for this purpose. Among them, Markov chains give a first framework for the modeling of population of cells [80], [57]. Piecewise deterministic processes are non diffusion processes also frequently used in the biological context [47], [56], [49]. Among Markov model, we developed strong expertise about processes derived from Brownian motion and Stochastic Differential Equations [72], [55]. For instance, knowledge about Brownian or random walk excursions [79], [71] helps to analyse genetic sequences and to develop inference about it. However, nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion of fractional fields. The basic aspects of these differential equations are now well understood, mainly thanks to the so-called rough paths tools [63], but also invoking the Russo-Vallois integration techniques [73]. The specific issue of Volterra equations driven by fractional Brownian motion, which is central for the subdiffusion within proteins problem, is addressed in [48]. Many generalizations (Gaussian or not) of this model have been recently proposed for some Gaussian locally self-similar fields, or for some non-Gaussian models [60], or for anisotropic models [44].

3.3. Estimation and control for stochastic processes

We develop inference about stochastic processes that we use for modeling. Control of stochastic processes is also a way to optimise administration (dose, frequency) of therapy.

There are many estimation techniques for diffusion processes or coefficients of fractional or multifractional Brownian motion according to a set of observations [59], [40], [46]. But, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. Our team has a good expertise about inference of the jump rate and the kernel of Piecewise Deterministic Markov Processes (PDMP) [37], [38], [36], [39]. However, there are many directions to go further into. For instance, previous works made the assumption of a complete observation of jumps and mode, that is unrealistic in practice. We tackle the problem of inference of "Hidden PDMP". As an example, in pharmacokinetics modeling inference, we want to take into account for presence of timing noise and identification from longitudinal data. We have expertise on this subjects [41], and we also used mixed models to estimate tumor growth [42].

We consider the control of stochastic processes within the framework of Markov Decision Processes [70] and their generalization known as multi-player stochastic games, with a particular focus on infinite-horizon problems. In this context, we are interested in the complexity analysis of standard algorithms, as well as the proposition and analysis of numerical approximate schemes for large problems in the spirit of [43]. Regarding complexity, a central topic of research is the analysis of the Policy Iteration algorithm, which has made significant progress in the last years [82], [69], [54], [78], but is still not fully understood. For large problems, we have a long experience of sensitivity analysis of approximate dynamic programming algorithms for Markov Decision Processes [76], [75], [77], [62], [74], and we currently investigate whether/how similar ideas may be adapted to multi-player stochastic games.

3.4. Algorithms and estimation for graph data

A graph data structure consists of a set of nodes, together with a set of pairs of these nodes called edges. This type of data is frequently used in biology because they provide a mathematical representation of many concepts such as biological structures and networks of relationships in a population. Some attention has recently been focused in the group on modeling and inference for graph data.

Network inference is the process of making inference about the link between two variables taking into account the information about other variables. [81] gives a very good introduction and many references about network inference and mining. Many methods are available to infer and test edges in Gaussian graphical models [81], [64], [52], [53]. However, when dealing with abundance data, because inflated zero data, we are far from gaussian assumption and we want to develop inference in this case.

Among graphs, trees play a special role because they offer a good model for many biological concepts, from RNA to phylogenetic trees through plant structures. Our research deals with several aspects of tree data. In particular, we work on statistical inference for this type of data under a given stochastic model. We also work on lossy compression of trees via directed acyclic graphs. These methods enable us to compute distances between tree data faster than from the original structures and with a high accuracy.

3.5. Regression and machine learning

Regression models and machine learning aim at inferring statistical links between a variable of interest and covariates. In biological study, it is always important to develop adapted learning methods both in the context of *standard* data and also for data of high dimension (with sometimes few observations) and very massive or online data.

Many methods are available to estimate conditional quantiles and test dependencies [68], [58]. Among them we have developed nonparametric estimation by local analysis via kernel methods [50], [51] and we want to study properties of this estimator in order to derive a measure of risk like confidence band and test. We study also many other regression models like survival analysis, spatio temporal models with covariates. Among the multiple regression models, we want to develop omnibus tests that examine several assumptions together.

Concerning the analysis of high dimensional data, our view on the topic relies on the *French data analysis school*, specifically on Factorial Analysis tools. In this context, stochastic approximation is an essential tool [61], which allows one to approximate eigenvectors in a stepwise manner [67], [65], [66]. BIGS aims at performing accurate classification or clustering by taking advantage of the possibility of updating the information "online" using stochastic approximation algorithms [45]. We focus on several incremental procedures for regression and data analysis like linear and logistic regressions and PCA (Principal Component Analysis).

We also focus on the biological context of high-throughput bioassays in which several hundreds or thousands of biological signals are measured for a posterior analysis. We have to account for the inter-individual variability within the modeling procedure. We aim at developing a new solution based on an ARX (Auto Regressive model with eXternal inputs) model structure using the EM (Expectation-Maximisation) algorithm for the estimation of the model parameters.

BIOCORE Project-Team

3. Research Program

3.1. Mathematical and computational methods

BIOCORE's action is centered on the mathematical modeling of biological systems, more particularly of artificial ecosystems, that have been built or strongly shaped by human. Indeed, the complexity of such systems where life plays a central role often makes them impossible to understand, control, or optimize without such a formalization. Our theoretical framework of choice for that purpose is Control Theory, whose central concept is "the system", described by state variables, with inputs (action on the system), and outputs (the available measurements on the system). In modeling the ecosystems that we consider, mainly through ordinary differential equations, the state variables are often population, substrate and/or food densities, whose evolution is influenced by the voluntary or involuntary actions of man (inputs and disturbances). The outputs will be some product that one can collect from this ecosystem (harvest, capture, production of a biochemical product, etc), or some measurements (number of individuals, concentrations, etc). Developing a model in biology is however not straightforward: the absence of rigorous laws as in physics, the presence of numerous populations and inputs in the ecosystems, most of them being irrelevant to the problem at hand, the uncertainties and noise in experiments or even in the biological interactions require the development of dedicated techniques to identify and validate the structure of models from data obtained by or with experimentalists.

Building a model is rarely an objective in itself. Once we have checked that it satisfies some biological constraints (eg. densities stay positive) and fitted its parameters to data (requiring tailor-made methods), we perform a mathematical analysis to check that its behavior is consistent with observations. Again, specific methods for this analysis need to be developed that take advantage of the structure of the model (eg. the interactions are monotone) and that take into account the strong uncertainty that is linked to life, so that qualitative, rather than quantitative, analysis is often the way to go.

In order to act on the system, which often is the purpose of our modeling approach, we then make use of two strong points of Control Theory: 1) the development of observers, that estimate the full internal state of the system from the measurements that we have, and 2) the design of a control law, that imposes to the system the behavior that we want to achieve, such as the regulation at a set point or optimization of its functioning. However, due to the peculiar structure and large uncertainties of our models, we need to develop specific methods. Since actual sensors can be quite costly or simply do not exist, a large part of the internal state often needs to be re-constructed from the measurements and one of the methods we developed consists in integrating the large uncertainties by assuming that some parameters or inputs belong to given intervals. We then developed robust observers that asymptotically estimate intervals for the state variables [83]. Using the directly measured variables and those that have been obtained through such, or other, observers, we then develop control methods that take advantage of the system structure (linked to competition or predation relationships between species in bioreactors or in the trophic networks created or modified by biological control).

3.2. A methodological approach to biology: from genes to ecosystems

One of the objectives of BIOCORE is to develop a methodology that leads to the integration of the different biological levels in our modeling approach: from the biochemical reactions to ecosystems. The regulatory pathways at the cellular level are at the basis of the behavior of the individual organism but, conversely, the external stresses perceived by the individual or population will also influence the intracellular pathways. In a modern "systems biology" view, the dynamics of the whole biosystem/ecosystem emerge from the interconnections among its components, cellular pathways/individual organisms/population. The different scales of size and time that exist at each level will also play an important role in the behavior of the biosystem/ecosystem. We intend to develop methods to understand the mechanisms at play at each level,

from cellular pathways to individual organisms and populations; we assess and model the interconnections and influence between two scale levels (eg., metabolic and genetic; individual organism and population); we explore the possible regulatory and control pathways between two levels; we aim at reducing the size of these large models, in order to isolate subsystems of the main players involved in specific dynamical behaviors.

We develop a theoretical approach of biology by simultaneously considering different levels of description and by linking them, either bottom up (scale transfer) or top down (model reduction). These approaches are used on modeling and analysis of the dynamics of populations of organisms; modeling and analysis of small artificial biological systems using methods of systems biology; control and design of artificial and synthetic biological systems, especially through the coupling of systems.

The goal of this multi-level approach is to be able to design or control the cell or individuals in order to optimize some production or behavior at higher level: for example, control the growth of microalgae via their genetic or metabolic networks, in order to optimize the production of lipids for bioenergy at the photobioreactor level.

BIOVISION Project-Team

3. Research Program

3.1. Introduction

The Biovision team has started on January 1st, 2016 and became an Equipe Projet Inria on August 1st, 2018 . It aims at developing fundamental research as well as technological developments along two axes.

3.1.1. Axis 1: High tech vision aid-systems for low-vision patients

Visual impairment, also known as vision loss, is a decreased ability to see to a degree that causes problems not fixable by usual means, such as glasses or lenses. Low-vision is a condition caused by eye disease, in which visual acuity is 20/70, meaning that the person is able to see, at 20 meters from a chart, what a normal person would see at 70 meters. Visual impairment affects some 285 million humans in the world, mostly in developed countries where this number is going to increase rapidly due to aging. 85% have low-vision or poorer. [0] There is a strong need to conceive new aid-systems to help these people in their daily living activities. Such systems already exist and can be divided into two categories according to their function. The first category concerns aids that translate visual information into alternative sensory information, such as touch or sound, called Sensory Substitution Devices (SSDs) [45], [40]. The second category concerns aids that adapt visual information to render it more visible to the patients, using scene processing methods and suitable devices. These are based on technological and algorithmic solutions that enhance salient scene characteristics [60], [56]. In Biovision team, we focus on this second category by targeting new vision aid-systems helping patients in their daily life, adapting to their own pathology.

We have strong contacts and collaborations with low-vision centers and associations in order to better understand low-vision patients needs, and have feedback on our prototypes aimed to be distributed to patients via transfer or company creation (startup). With the fast-growing number of incurable eye diseases, crucial steps must be taken to increase visual accessibility by:

- Designing solutions for earlier and more decisive detection of visual pathologies,
- Developing efficient rehabilitation protocols, and,
- Designing innovative vision-aid systems to empower patients with improved perceptual capacities.

To do this, we need to work in synergy with patients to assess their needs, understand their pathologies at a perceptual level and design personalized solutions to create change and adoption. This will require developing state-of-the-art methods in computer science, necessitating skills from many areas such as artificial intelligence, virtual and augmented reality, human-machine interface, multimedia systems, etc. By doing so, we will leverage new technologies to offer life-changing solutions for people with visual impairment [12], [15].

3.1.2. Axis 2: Human vision understanding through joint experimental and modeling studies, for normal and distrophic retinas

A holistic point of view is emerging in neuroscience where one can observe simultaneously how vision works at different levels of the hierarchy in the visual system. Multiple scales functional analysis and connectomics are also exploding in brain science, and studies of visual systems are upfront on this fast move. These integrated studies call for new classes of theoretical and integrated models where the goal is the modeling of visual functions such as motion integration.

[0] Source: VisionAware

In Biovision we contribute to a better understanding of the visual system with those main goals:

1. Proposing simplified mathematical models characterizing how the retina converts a visual scene into **spike population coding**, in normal and under specific pathological conditions.
2. Designing biophysical models allowing to better understand the **multiscale dynamics** of the retina, from dynamics of individual cells to their collective activity, and how changes in biophysical parameters (development, pharmacology, pathology) impacts this dynamics.
3. Elaborating an **integrated mathematical and numerical model** of the visual stream, with a focus on motion integration, from retina to early visual cortex (V1).
4. Developing a **simulation platform** emulating the retinal response to visual and prosthetic simulations, enabling us to test hypotheses about the functioning of the early visual system, in normal, pharmacological or pathological conditions.

Finally, although this is not the main goal of our team, two other natural avenues of our research are (i) to develop novel synergistic solutions to solve computer vision tasks based on bio-inspired mechanisms [7]; (ii) collaborate with neuroscientists and neuronal modellers to address mathematical problems outside the scope of the retina or the early visual system.

3.2. Scientific methodology

In this section we briefly describe the scientific methods we use to achieve our research goals.

3.2.1. Adaptive image enhancement

Image enhancement is a natural type of image processing method to help low-vision people better understand visual scenes. An impressive number of techniques have been developed in the fields of computer vision and computer graphics to manipulate image content for a variety of applications. Some of these methods have a direct interest in the design of vision aid-systems. Only a few of them have been carefully evaluated with patients [36], [49], [50], [41], [37]. Our objective is to further exploit and evaluate them with patients, considering dedicated use-cases, using virtual and augmented reality technology (Sec. 3.2.2). We consider not only classical brightness manipulations (e.g., equalization, gamma correction, tone mapping, edge enhancement, image decomposition and cartoonization) but also more sophisticated approaches which can change the geometric information of the scene to highlight the most relevant information (e.g., scene retargeting and seam carving). In addition, we investigate how image enhancements could be adapted to patients needs by relating tuning parameters to the patient pathology.

3.2.2. Virtual, mixed and augmented reality

Virtual, mixed and augmented reality technology (VR/MR/AR) is based on the idea of combining digital worlds with physical realities in different ways. It encompasses a wide spectrum of hardware. It is our conviction that this technology will play a major role in the domain of low-vision. Not only can this technology be useful to design novel vision aid-systems and rehabilitation programs, but also it has the potential to revolutionize how we study the behaviour of low-vision people (controlled condition, free head, eye tracking, possibilities for large scale studies). These projects require a constant interaction with psychophysicists and ophtalmologists so as to design our solutions based on patients needs and capabilities.

3.2.3. Biophysical modeling

Modeling in neuroscience has to cope with several competing objectives. On one hand, describing the biological realm as close as possible, and, on the other hand, providing tractable equations at least at the descriptive level (simulation, qualitative description) and, when possible, at the mathematical level (i.e., affording a rigorous description). These objectives are rarely achieved simultaneously and most of the time one has to make compromises. In the Biovision team we adopt the point of view of a physicist: try to capture the phenomenological description of a biophysical mechanism, removing irrelevant details in the description, and try to have a qualitative description of equations behaviour at least at the numerical simulation level, and, when possible, obtain analytical results. We insist on the quality of the model in predicting and proposing new experiments. This requires a constant interaction with neuroscientists so as to keep the model on the tracks, warning of too crude approximation, still trying to construct equations from canonical principles [1], [2], [6].

3.2.4. Methods from theoretical physics

Biophysical models mainly consist of differential equations (ODEs or PDEs) or integro-differential equations (neural fields). We study them using dynamical systems and bifurcation theory as well as techniques coming from nonlinear physics (amplitude equations, stability analysis, Lyapunov spectrum, correlation analysis, multi-scales methods) [23].

For the study of large scale populations (e.g., when studying population coding) we use methods coming from statistical physics. This branch of physics gave birth to mean-field methods as well statistical methods for large population analysis. We use both of them. Mean-field methods are applied for large scale activity in the retina and in the cortex [4], [8], [39].

For the study of retina population coding we use the so-called Gibbs distribution, initially introduced by Boltzmann and Gibbs. This concept includes, but *is not limited to*, maximum entropy models [55] used by numerous authors in the context of the retina (see, e.g., [57], [59], [52], [51], [61]). These papers were restricted to a statistical description without memory neither causality: the time correlations between successive times is not considered. However, maximum entropy extends to spatio-temporal correlations as we have shown in, e.g., [2] [62], [43]. In this context, we study how the retina respond to transient stimuli (moving objects), i.e. how spatio-temporal correlations are modified when a moving object crosses the receptive fields of ganglion cells, taking into account the lateral connectivity due to amacrine cells [42], [20], [11], [21].

CAMIN Project-Team

3. Research Program

3.1. Exploration and understanding of the origins and control of movement

One of CAMIN's areas of expertise is **motion measurement, observation and modeling** in the context of **sensorimotor deficiencies**. The team has the capacity to design advanced protocols to explore motor control mechanisms in more or less invasive conditions in both animal and human.
Human movement can be assessed by several noninvasive means, from motion observation (MOCAP, IMU) to electrophysiological measurements (afferent ENG, EMG, see below). Our general approach is to develop solutions that are realistic in terms of clinical or home use by clinical staff and/or patients for diagnosis and assessment purposes. In doing so, we try to gain a better understanding of motor control mechanisms, including deficient ones, which in turn will give us greater insight into the basics of human motor control. Our ultimate goal is to optimally match a neuroprosthesis to the targeted sensorimotor deficiency.
The team is involved in research projects including:

- Peripheral nervous system (PNS) exploration, modeling and electrophysiology techniques
  Electroneurography (ENG) and electromyography (EMG) signals inform about neural and muscular activities. The team investigates both natural and evoked ENG/EMG through advanced and dedicated signal processing methods. Evoked responses to ES are very precious information for understanding neurophysiological mechanisms, as both the input (ES) and the output (evoked EMG/ENG) are controlled. CAMIN has the expertise to perform animal experiments (rabbits, rats, earthworms and big animals with partners), design hardware and software setups to stimulate and record in harsh conditions, process signals, analyze results and develop models of the observed mechanisms. Experimental surgery is mandatory in our research prior to invasive interventions in humans. It allows us to validate our protocols from theoretical, practical and technical aspects.
- Central nervous system (CNS) exploration
  Stimulating the CNS directly instead of nerves allows activation of the neural networks responsible for generating functions. Once again, if selectivity is achieved the number of implanted electrodes and cables would be reduced, as would the energy demand. We have investigated **spinal electrical stimulation** in animals (pigs) for urinary track and lower limb function management. This work is very important in terms of both future applications and the increase in knowledge about spinal circuitry. The challenges are technical, experimental and theoretical, and the preliminary results have enabled us to test some selectivity modalities through matrix electrode stimulation. This research area will be further intensified in the future as one of ways to improve neuroprosthetic solutions.
  We intend to gain a better understanding of the electrophysiological effects of DES through electroencephalographic (EEG) and electrocorticographic (ECoG) recordings in order to optimize anatomo-functional brain mapping, better understand brain dynamics and plasticity, and improve surgical planning, rehabilitation, and the quality of life of patients.
- Muscle models and fatigue exploration
  Muscle fatigue is one of the major limitations in all FES studies. Simply, the muscle torque varies over time even when the same stimulation pattern is applied. As there is also muscle recovery when there is a rest between stimulations, modeling the fatigue is almost an impossible task. Therefore, it is essential to monitor the muscle state and assess the expected muscle response by FES to improve the current FES system in the direction of greater adaptive force/torque control in the presence of muscle fatigue.
- Movement interpretation

We intend to develop ambulatory solutions to allow ecological observation. We have extensively investigated the possibility of using inertial measurement units (IMUs) within body area networks to observe movement and assess posture and gait variables. We have also proposed extracting gait parameters like stride length and foot-ground clearance for evaluation and diagnosis purposes.

3.2. Movement assistance and/or restoration

The challenges in movement restoration are: (i) improving nerve/muscle stimulation modalities and efficiency and (ii) global management of the function that is being restored in interaction with the rest of the body under voluntary control. For this, both local (muscle) and global (function) controls have to be considered.
Online modulation of ES parameters in the context of lower limb functional assistance requires the availability of information about the ongoing movement. Different levels of complexity can be considered, going from simple open-loop to complex control laws (Figure 2).

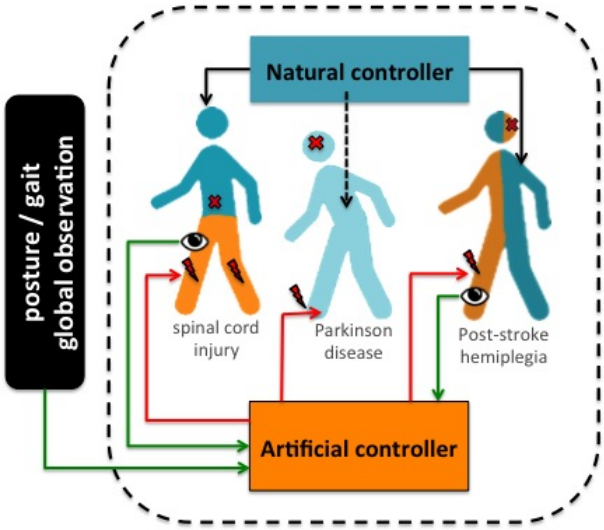


Figure 2. FES assistance should take into account the coexistence of artificial and natural controllers. Artificial controllers should integrate both global (posture/gait) and local (limb/joint) observations.

Real-time adaptation of the stimulation patterns is an important challenge in most of the clinical applications we consider. The modulation of ES parameters to adapt to the occurrence of muscular fatigue or to environement changes needs for advanced adaptive controllers based on sensory information. A special care in minimizing the number of sensors and their impact on patient motion should be taken.

3.3. On-going clinical protocols

One specificity of CAMIN team is to be involved in Clinical protocols. At the moment we are involved in the following protocols:

- CYCLOSEF: Training spinal cord injured people pedaling a tricycle assisted by electric stimulation of sublesional muscles: case study - Protocol RCB 2019-A00808-49. CRF La Châtaigneraie.
- AGILIS - Functional evaluation of the recovery of prehension in quadriplegics by implanted neural stimulation - Protocol RCB 2019-A02037-50. APHP (Paris)

- E-PREHENSTROKE - Evaluation of optimal piloting modalities and their impact on the grasping capacity in Functional Electrical Stimulation of Finger Extensor Muscles in the Hemiplegic Patient in Chronic Phase - Protocol RCB 2018-A02144-51. CHU Nîmes.
- PBREATHLOOP - Recording tracheal sounds for the purpose of developing a breath control algorithm - Protocol RCB 2019-A01813-54. ADOREPS
- Variability and evolution of the single fiber potentials of a spastic muscle treated with botulinum toxin - Protocol RCB 2019-A01863-52A. CHU Nîmes
- Pilot study: measurement of evoked potentials in electroencephalography and electrocorticography by electrical stimulation of the brain during awake neuro-surgery of low-grade infiltrating gliomas - Protocol RCB 2014-A00056-43. CHU Montpellier

CAPSID Project-Team

3. Research Program

3.1. Classifying and Mining Protein Structures and Protein Interactions

3.1.1. Context

The scientific discovery process is very often based on cycles of measurement, classification, and generalisation. It is easy to argue that this is especially true in the biological sciences. The proteins that exist today represent the molecular product of some three billion years of evolution. Therefore, comparing protein sequences and structures is important for understanding their functional and evolutionary relationships [67], [48]. There is now overwhelming evidence that all living organisms and many biological processes share a common ancestry in the tree of life. Historically, much of bioinformatics research has focused on developing mathematical and statistical algorithms to process, analyse, annotate, and compare protein and DNA sequences because such sequences represent the primary form of information in biological systems. However, there is growing evidence that structure-based methods can help to predict networks of protein-protein interactions (PPIs) with greater accuracy than those which do not use structural evidence [52], [70]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [39].

3.1.2. Formalising and Exploiting Domain Knowledge

Concerning protein structure classification, we aim to explore novel classification paradigms to circumvent the problems encountered with existing hierarchical classifications of protein folds and domains. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [43], but instead using Kpax domain-domain similarity matrices. A non-trivial issue with fuzzy clustering is how to handle similarity rather than mathematical distance matrices, and how to find the optimal number of clusters, especially when using a non-Euclidean similarity measure. We will adapt the algorithms and the calculation of quality indices to the Kpax similarity measure. More fundamentally, it will be necessary to integrate this classification step in the more general process leading from data to knowledge called Knowledge Discovery in Databases (KDD) [46].

Another example where domain knowledge can be useful is during result interpretation: several sources of knowledge have to be used to explicitly characterise each cluster and to help decide its validity. Thus, it will be useful to be able to express data models, patterns, and rules in a common formalism using a defined vocabulary for concepts and relationships. Existing approaches such as the Molecular Interaction (MI) format [49] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [58]. A different point of view is represented in the Interaction Network Ontology (INO), a community-driven ontology that aims to standardise and integrate data on interaction networks and to support computer-assisted reasoning [71]. However, this ontology does not integrate basic 3D concepts and structural relationships. Therefore, extending such formalisms and symbolic relationships will be beneficial, if not essential, when classifying the 3D shapes of proteins at the domain family level.

Domain family classification is also relevant for studying domain-domain interactions (DDI). Our previous work on Knowledge-Based Docking (KBDOCK, [3], [5] will be updated and extended using newly published DDIs. Methods for inferring new DDIs from existing protein-protein interactions (PPIs) will be developped. Efforts should be made for validating such inferred DDIs so that they can be used to enrich DDI classification and predict new PPIs.

In parallel, we also intend to design algorithms for leveraging information embedded in biological knowledge graphs (also known as complex networks). Knowledge graphs mostly represent PPIs, integrated with various properties attached to proteins, such as pathways, drug binding or relation with diseases. Setting up similarity measures for proteins in a knowledge graph is a difficult challenge. Our objective is to extract useful knowledge from such graphs in order to better understand and highlight the role of multi-component assemblies in various types of cell or organisms. Ultimately, knowledge graphs can be used to model and simulate the functioning of such molecular machinery in the context of the living cell, under physiological or pathological conditions.

3.1.3. Function Annotation in large protein graphs

Knowledge of the functional properties of proteins can shed considerable light on how they might interact. However, huge numbers of protein sequences in public databases such as UniProt/TrEMBL lack any functional annotation, and the functional annotation of such sequences is a highly challenging problem. We are developing graph-based and machine learning techniques to annotate automatically the available unannotated sequences with functional properties such as EC numbers and Gene Ontology (GO) terms (note that these terms are organized hierarchically allowing generalization/specialization reasoning). The idea is to transfer annotations from expert-reviewed sequences present in the UniProt/SwissProt database (about 560 thousands entries) to unreviewed sequences present in the UniProt/TrEMBL database (about 80% of 180 millions entries). For this, we have to learn from the UniProt/SwissProt database how to compute the similarity of proteins sharing identical or similar functional annotations. Various similarity measures can be tested using cross-validation approches in the UniProt/SwissProt database. For instance, we can use primary sequence or domain signature similarities. More complex similarities can be computed with graph-embedding techniques.

This work is in progress with Bishnu Sarker's PhD project and a first approach called GrAPFI (Graph-based Automatic Protein Function Inference) was presented at conferences in 2018 [11], [12].

3.2. Integrative Multi-Component Assembly and Modeling

3.2.1. Context

At the molecular level, each PPI is embodied by a physical 3D protein-protein interface. Therefore, if the 3D structures of a pair of interacting proteins are known, it should in principle be possible for a docking algorithm to use this knowledge to predict the structure of the complex. However, modeling protein flexibility accurately during docking is very computationally expensive. This is due to the very large number of internal degrees of freedom in each protein, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead, most protein docking algorithms use fast heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2. Polar Fourier Docking Correlations

In our *Hex* protein docking program [60], the shape of a protein molecule is represented using polar Fourier series expansions of the form

$$\sigma(\underline{x}) = \sum_{nlm} a_{nlm} R_{nl}(r) y_{lm}(\theta, \phi), \tag{56}$$

where $\sigma(\underline{x})$ is a 3D shape-density function, a_{nlm} are the expansion coefficients, $R_{nl}(r)$ are orthonormal Gauss-Laguerre polynomials and $y_{lm}(\theta, \phi)$ are the real spherical harmonics. The electrostatic potential, $\phi(\underline{x})$, and charge density, $\rho(\underline{x})$, of a protein may be represented using similar expansions. Such representations allow the *in vacuo* electrostatic interaction energy between two proteins, A and B, to be calculated as [51]

$$E = \frac{1}{2} \int \phi_A(\underline{x}) \rho_B(\underline{x}) \mathrm{d}\underline{x} + \frac{1}{2} \int \phi_B(\underline{x}) \rho_A(\underline{x}) \mathrm{d}\underline{x}. \tag{57}$$

This equation demonstrates using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [61]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10], [7]. Consequently, we are continuing to explore new ways to exploit the polar Fourier approach.

3.2.3. Assembling Symmetrical Protein Complexes

Although protein-protein docking algorithms are improving [62], [53], it still remains challenging to produce a high resolution 3D model of a protein complex using *ab initio* techniques. This is mainly due to the problem of structural flexibility described above. However, with the aid of even just one simple constraint on the docking search space, the quality of docking predictions can improve considerably [10], [61]. In particular, many protein complexes involve symmetric arrangements of one or more sub-units, and the presence of symmetry may be exploited to reduce the search space considerably [38], [59], [66]. For example, using our operator notation (in which $\widehat{R}$ and $\widehat{T}$ represent 3D rotation and translation operators, respectively), we have developed an algorithm which can generate and score candidate docking orientations for monomers that assemble into cyclic (C_n) multimers using 3D integrals of the form

$$E_{AB}(y,\alpha,\beta,\gamma) = \int \left[\widehat{T}(0,y,0)\widehat{R}(\alpha,\beta,\gamma)\phi_A(\underline{x})\right] \times \left[\widehat{R}(0,0,\omega_n)\widehat{T}(0,y,0)\widehat{R}(\alpha,\beta,\gamma)\rho_B(\underline{x})\right] \mathrm{d}\underline{x}, \quad (58)$$

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n-fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n-mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n, D_n, T, O, I). This approach was published in 2016 [8], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [45]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to build and refine candidate solutions using a more accurate coarse-grained (CG) force-field scoring function.

3.2.4. Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use "coarse-grained" (CG) normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [57], [44], [54], [55]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [68], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [4].

In the last few years, CG force-field models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [37]. Typically, a CG force-field representation replaces the atoms in each amino acid with from 2 to 4 "pseudo-atoms", and it assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [65]. Furthermore, this kind of CG model effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [50]. We are currently developing a CG scoring function for fast protein-protein docking and multi-component assembly. This work is part of

the PhD project of Maria-Elisa Ruiz-Echartea [19], [64]. Beyond this PhD project, the CG scoring function will be exploited in all our docing projects, especially for RNA-Protein docking (see below).

3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recent developments in cryo-EM instruments and technologies, it is now feasible to capture low resolution images of very large macromolecular machines. However, while such developments offer the intriguing prospect of being able to trap biological systems in unprecedented levels of detail, there will also come with an increasing need to analyse, annotate, and interpret the enormous volumes of data that will soon flow from the latest instruments. In particular, a new challenge that is emerging is how to fit previously solved high resolution protein structures into low resolution cryo-EM density maps. However, the problem here is that large molecular machines will have multiple sub-components, some of which will be unknown, and many of which will fit each part of the map almost equally well. Thus, the general problem of building high resolution 3D models from cryo-EM data is like building a complex 3D jigsaw puzzle in which several pieces may be unknown or missing, and none of which will fit perfectly. We wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function [42], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space. This work has made some progress during the PhD project of Maria Elisa Ruiz Echartea but still requires further efforts.

3.2.6. Protein-Nucleic Acids Interactions

As well as playing an essential role in the translation of DNA into proteins, RNA molecules carry out many other essential biological functions in cells, often through their interactions with proteins. A critical challenge in modelling such interactions computationally is that the RNA is often highly flexible, especially in single-stranded (ssRNA) regions of its structure. These flexible regions are often very important because it is through their flexibility that the RNA can adjust its 3D conformation in order to bind to a protein surface. However, conventional protein-protein docking algorithms generally assume that the 3D structures to be docked are rigid, and so are not suitable for modeling protein-RNA interactions. There is therefore much interest in developing protein-RNA docking algorithms which can take RNA flexibility into account. This research topic has been initiated with the recruitement of Isaure Chauvot de Beauchêne in 2016 and is becoming a major activity in the team. A novel flexible docking algorithm is currently under development in the team. It first docks small fragments of ssRNA (typically three nucleotides at a time) onto a protein surface, and then combinatorially reassembles those fragments in order to recover a contiguous ssRNA structure on the protein surface [41], [40].

As the correctness of the initial docking of the fragments settles an upper limit to the correctness of the full model, we are now focusing on improving that step. A key component of our docking tool is the energy function of the protein - fragment interactions, that is used both to drive the sampling (positioning of the fragments) by minimization and to discriminate the correct final positions from decoys (i.e. false positives). We are developing a new knowledge-based energy function that will be learnt by machine-learning methods from public structural data on ssRNA-protein complexes.

In the future, we will improve the combinatorial algorithm used for reassembling the docked fragments using experimental constraints and machine-learning approaches.

CARMEN Project-Team

3. Research Program

3.1. Complex models for the propagation of cardiac action potentials

The contraction of the heart is coordinated by a complex electrical activation process which relies on about a million ion channels, pumps, and exchangers of various kinds in the membrane of each cardiac cell. Their interaction results in a periodic change in transmembrane potential called an action potential. Action potentials in the cardiac muscle propagate rapidly from cell to cell, synchronizing the contraction of the entire muscle to achieve an efficient pump function. The spatio-temporal pattern of this propagation is related both to the function of the cellular membrane and to the structural organization of the cells into tissues. Cardiac arrythmias originate from malfunctions in this process. The field of cardiac electrophysiology studies the multiscale organization of the cardiac activation process from the subcellular scale up to the scale of the body. It relates the molecular processes in the cell membranes to the propagation process and to measurable signals in the heart and to the electrocardiogram, an electrical signal on the torso surface.

Several improvements of current models of the propagation of the action potential are being developed in the Carmen team, based on previous work [56] and on the data available at IHU LIRYC:

- Enrichment of the current monodomain and bidomain models [56], [67] by accounting for structural heterogeneities of the tissue at an intermediate scale. Here we focus on multiscale analysis techniques applied to the various high-resolution structural data available at the LIRYC.
- Coupling of the tissues from the different cardiac compartments and conduction systems. Here, we develop models that couple 1D, 2D and 3D phenomena described by reaction-diffusion PDEs.

These models are essential to improve our in-depth understanding of cardiac electrical dysfunction. To this aim, we use high-performance computing techniques in order to numerically explore the complexity of these models.

We use these model codes for applied studies in two important areas of cardiac electrophysiology: atrial fibrillation [60] and sudden-cardiac-death (SCD) syndromes [7], [6] [64]. This work is performed in collaboration with several physiologists and clinicians both at IHU Liryc and abroad.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the cardiac electric signals is based on accurate reconstruction of the patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developed. This problem involves solving inverse problems that cannot be addressed with the more compex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally, the whole cardiac electrical activity, from high-density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough. Although widely studied recently, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not adequate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface. Our aim is to

- study in depth the dependance of this inverse problem on inhomogeneities in the torso, conductivity values, the geometry, electrode positions, etc., and
- improve the solution to the inverse problem by using new regularization strategies, factorization of boundary value problems, and the theory of optimal control.

Of course we will use our models as a basis to regularize these inverse problems. We will consider the following strategies:

- using complete propagation models in the inverse problem, like the bidomain equations, for instance in order to localize electrical sources;
- constructing families of reduced-order models using e.g. statistical learning techniques, which would accurately represent some families of well-identified pathologies; and
- constructing simple models of the propagation of the activation front, based on eikonal or level-set equations, but which would incorporate the representation of complex activation patterns.

Additionaly, we will need to develop numerical techniques dedicated to our simplified eikonal/level-set equations.

3.3. Numerical techniques

We want our numerical simulations to be efficient, accurate, and reliable with respect to the needs of the medical community. Based on previous work on solving the monodomain and bidomain equations [4], [5], [8], [1], we will focus on

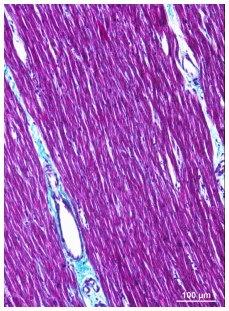
- High-order numerical techniques with respect to the variables with physiological meaning, like velocity, AP duration and restitution properties.
- Efficient, dedicated preconditioning techniques coupled with parallel computing.

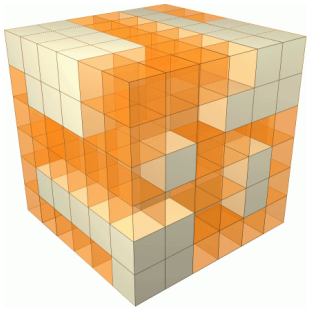
Existing simulation tools used in our team rely, among others, on mixtures of explicit and implicit integration methods for ODEs, hybrid MPI-OpenMP parallellization, algebraic multigrid preconditioning, and Krylov solvers. New developments include high-order explicit integration methods and task-based dynamic parallelism.

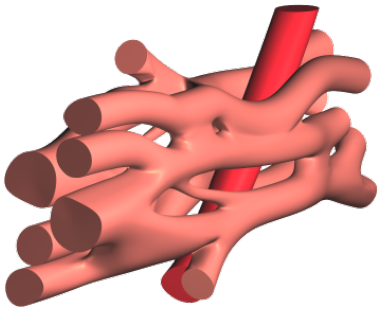
3.4. Cardiac Electrophysiology at the Microscopic Scale

Numerical models of whole-heart physiology are based on the approximation of a perfect muscle using homogenisation methods. However, due to aging and cardiomyopathies, the cellular structure of the tissue changes. These modifications can give rise to life-threatening arrhythmias. For our research on this subject and with cardiologists of the IHU LIRYC Bordeaux, we aim to design and implement models that describe the strong heterogeneity of the tissue at the cellular level and to numerically explore the mechanisms of these diseases.

The literature on this type of model is still very limited [74]. Existing models are two-dimensional [65] or limited to idealized geometries, and use a linear (purely resistive) behaviour of the gap-juction channels that connect the cells. We propose a three-dimensional approach using realistic cellular geometry (figure 1), nonlinear gap-junction behaviour, and a numerical approach that can scale to hundreds of cells while maintaining a sub-micrometer spatial resolution (10 to 100 times smaller than the size of a cardiomyocyte) [52], [51], [49]. P-E. Bécue defended his PhD thesis on this topic in December 2018.







A ***B*** ***C***

Figure 1. ***A:*** *The cardiac muscle consists of a branching network of elongated muscle cells, interspersed with other structures. Sheets of connective tissue (blue) can grow between the muscle cells and become pathogenic.* ***B:*** *Current models can only represent such alterations in a coarse way by replacing model elements with different types; each cube in this illustration would represent hundreds of cells.* ***C:*** *This hand-crafted example illustrates the type of geometric model we are experimenting with. Each cell is here represented by hundreds of elements.*

CASTOR Project-Team

3. Research Program

3.1. Plasma Physics

Participants: Jacques Blum, Cédric Boulbe, Blaise Faugeras, Hervé Guillard, Holger Heumann, Sebastian Minjeaud, Boniface Nkonga, Richard Pasquetti, Afeintou Sangam.

The main reseach topics are:

1. Modelling and analysis
   - Fluid closure in plasma
   - Turbulence
   - Plasma anisotropy type instabilities
   - Free boundary equilibrium (FBE)
   - Coupling FBE – Transport
2. Numerical methods and simulations
   - High order methods
   - Curvilinear coordinate systems
   - Equilibrium simulation
   - Pressure correction scheme
   - Anisotropy
   - Solving methods and parallelism
3. Identification and control
   - Inverse problem: Equilibrium reconstruction
   - Open loop control
4. Applications
   - MHD instabilities : Edge-Localized Modes (ELMs)
   - Edge plasma turbulence
   - Optimization of scenarii

COFFEE Project-Team

3. Research Program

3.1. Research Program

Mathematical modeling and computer simulation are among the main research tools for environmental management, risks evaluation and sustainable development policy. Many aspects of the computer codes as well as the PDEs systems on which these codes are based can be considered as questionable regarding the established standards of applied mathematical modeling and numerical analysis. This is due to the intricate multiscale nature and tremendous complexity of those phenomena that require to set up new and appropriate tools. Our research group aims to contribute to bridging the gap by developing advanced abstract mathematical models as well as related computational techniques.

The scientific basis of the proposal is two–fold. On the one hand, the project is "technically–driven": it has a strong content of mathematical analysis and design of general methodology tools. On the other hand, the project is also "application–driven": we have identified a set of relevant problems motivated by environmental issues, which share, sometimes in a unexpected fashion, many common features. The proposal is precisely based on the conviction that these subjects can mutually cross-fertilize and that they will both be a source of general technical developments, and a relevant way to demonstrate the skills of the methods we wish to design.

To be more specific:

- We consider evolution problems describing highly heterogeneous flows (with different phases or with high density ratio). In turn, we are led to deal with non linear systems of PDEs of convection and/or convection–diffusion type.
- The nature of the coupling between the equations can be two–fold, which leads to different difficulties, both in terms of analysis and conception of numerical methods. For instance, the system can couple several equations of different types (elliptic/parabolic, parabolic/hyperbolic, parabolic or elliptic with algebraic constraints, parabolic with degenerate coefficients....). Furthermore, the unknowns can depend on different sets of variables, a typical example being the fluid/kinetic models for particulate flows. In turn, the simulation cannot use a single numerical approach to treat all the equations. Instead, hybrid methods have to be designed which raise the question of fitting them in an appropriate way, both in terms of consistency of the discretization and in terms of stability of the whole computation. For the problems under consideration, the coupling can also arises through interface conditions. It naturally occurs when the physical conditions are highly different in subdomains of the physical domain in which the flows takes place. Hence interface conditions are intended to describe the exchange (of mass, energy...) between the domains. Again it gives rise to rather unexplored mathematical questions, and for numerics it yields the question of defining a suitable matching at the discrete level, that is requested to preserve the properties of the continuous model.
- By nature the problems we wish to consider involve many different scales (of time or length basically). It raises two families of mathematical questions. In terms of numerical schemes, the multiscale feature induces the presence of stiff terms within the equations, which naturally leads to stability issues. A clear understanding of scale separation helps in designing efficient methods, based on suitable splitting techniques for instance. On the other hand asymptotic arguments can be used to derive hierarchy of models and to identify physical regimes in which a reduced set of equations can be used.

We can distinguish the following fields of expertise

- Numerical Analysis: Finite Volume Schemes, Well-Balanced and Asymptotic-Preserving Methods
  - Finite Volume Schemes for Diffusion Equations and Viscous Flows
  - Finite Volume Schemes for Conservation Laws
  - Well-Balanced and Asymptotic-Preserving Methods
  - Domain Decomposition Methods
- Modeling and Analysis of PDEs
  - Kinetic equations and hyperbolic systems
  - PDEs in random media
  - Interface problems

COMMEDIA Project-Team

3. Research Program

3.1. Multi-physics modeling and simulation

The research activity in terms of modeling and simulation (i.e., the so-called forward problem) is driven by two application domains related to the cardiovascular and the respiratory systems.

3.1.1. Cardiovascular hemodynamics

We distinguish between *cardiac hemodynamics* (blood flow inside the four chambers of the heart) and *vascular hemodynamics* (blood flow in the vessels of the body).

Cardiac hemodynamics. The numerical simulation of cardiac hemodynamics presents many difficulties. We can mention, for instance, the large deformation of the cardiac chambers and the complex fluid-structure interaction (FSI) phenomena between blood, the valves and the myocardium. Blood flow can be described by the incompressible Navier-Stokes equations which have to be coupled with a bio-physical model of the myocardium electro-mechanics and a mechanical model of the valves. The coupling between the fluid and the solid media is enforced by kinematic and dynamic coupling conditions, which guarantee the continuity of velocity and stresses across the interface. In spite of the significant advances achieved since the beginning of this century (see, e.g., [61], [69], [60], [63], [53]), the simulation of all the fluid-structure interaction phenomena involved in the heart hemodynamics remains a complex and challenging problem.

Heart valves are definitely a bottleneck of the problem, particularly due to their fast dynamics and the contact phenomena at high pressure-drops. Computational cost is recognized as one of the key difficulties, related to the efficiency of the FSI coupling method and the robustness of the contact algorithm. Furthermore, the numerical discretization of these coupled systems requires to deal with unfitted fluid and solid meshes, which are known to complicate the accuracy and/or the robustness of the numerical approximations (see Section 3.3.2 below).

The ultimate goal of the proposed research activity is the simulation of the complete fluid-structure-contact interaction phenomena involved within the heart. Most of this work will be carried out in close collaboration with the M3DISIM project-team, which has a wide expertise on the modeling, simulation and estimation of myocardium electro-mechanics. We will also consider simplified approaches for cardiac hemodynamics (see, e.g., [34], [48], [51]). The objective is to develop mathematically sound models of reduced valve dynamics with the purpose of enhancing the description of the pressure dynamics right after the opening/closing of the valve (traditional models yield spurious pressure oscillations).

Vascular hemodynamics. The modeling and simulation of vascular hemodynamics in large vessels has been one of the core research topics of some members of COMMEDIA, notably as regards the fluid-structure interaction phenomena. Here we propose to investigate the modeling of pathological scenarios, such as the hemorrhage phenomena in smaller vessels. Modeling of hemorrhage is motivated by the medical constatation that, after a primary vessel wall rupture, secondary vessel wall ruptures are observed. Biologists postulate that the mechanical explanation of this phenomena might be in the change of applied stress due to blood bleeding. We propose to model and simulate the underlying coupled system, blood vessel flow through the external tissue, to estimate the effect of the subsequent stress variation.

3.1.2. Respiratory flows

The motivation of the proposed research activities is to develop a hierarchy of easily parametrizable models allowing to describe and efficiently simulate the physical, mechanical and biological phenomena related to human respiration, namely, ventilation, particle deposition, gas diffusion and coupling with the circulatory system.

Ventilation. The current modeling approaches (either 3D–0D coupled models where the 3D Navier-Stokes equations are solved in truncated geometries of the bronchial tree with appropriate lumped boundary conditions, or 0D–3D coupled models where the lung parenchyma is described by a 3D elastic media irrigated by a simplified bronchial tree) provide satisfactory results in the case of mechanical ventilation or normal breathing. Realistic volume-flow phase portraits can also be simulated in the case of forced expiration (see [36], [45], [66]), but the magnitude of the corresponding pressure is not physiological. The current models must be enriched since they do not yet correctly describe all the physiological phenomena at play. We hence propose to extend the 0D–3D (bronchial tree–parenchyma) model developed in the team, by considering a non-linear, viscoelastic and possibly poro-elastic description of the parenchyma with appropriate boundary conditions that describe ribs and adjacent organs and taking into account an appropriate resistive model.

So far, the motion of the trachea and proximal bronchi has been neglected in the ventilation models (see, e.g., [67]). These features can be critical for the modeling of pathologic phenomena such as sleep apnea and occlusion of the airways. This would be a long-term goal where fluid-structure interaction and the possible contact phenomena will be taken into account, as in the simulation of cardiac hemodynamics (see Section 3.1.1).

Aerosol and gas diffusion. The dynamics of aerosols in the lung have been widely studied from the mathematical modeling standpoint. They can be described by models at different scales: the microscopic one for which each particle is described individually, the mesoscopic (or kinetic) one for which a density of probability is considered, or the macroscopic one where reaction-diffusion equations describing the behavior of the constituant concentration are considered. The objective of COMMEDIA will mainly be to develop the kinetic approach that allows a precise description of the deposition area at controlled computational costs. Part of this study could be done in collaboration with colleagues from the Research Center for Respiratory Diseases at Inserm Tours (UMR1100).

The macroscopic description is also appropriate for the diffusion of gases (oxygen and carbon dioxide) in the bronchial tree (see [62]). Regarding the influence of the carrier gas, if the patient inhales a different mixture of air such as a Helium-Oxygen mixture, the diffusion mechanisms could be modified. In this context, the goal is to evaluate if the cross-diffusion (and thus the carrier gas) modifies the quantities of oxygen diffused. Part of this work will be carried out in collaboration with members of the LJLL and of the MAP5.

As a long term goal, we propose to investigate the coupling of these models to models of diffusion in the blood or to perfusion models of the parenchyma, and thus, have access thanks to numerical simulations to new indices of ventilation efficiency (such as dissolved oxygen levels), depending on the pathology considered or the resting or exercise condition of the patient.

3.2. Simulation with data interaction

The second research axis of COMMEDIA is devoted to the interaction of numerical simulations with measured data. Several research directions related to two specific applications are described below: blood flows and cardiac electrophysiology, for which the mathematical models have been validated against experimental data. This list is not exhaustive and additional problems (related to cardiac and respiratory flows) shall be considered depending on the degree of maturity of the developed models.

3.2.1. Fluid flow reconstruction from medical imaging

A first problem which is currently under study at COMMEDIA is the reconstruction of the flow state from Doppler ultrasound measurements. This is a cheap and largely available imaging modality where the measure can be interpreted as the average on a voxel of the velocity along the direction of the ultrasound beam. The goal is to perform a full-state estimation in a time compatible with a realistic application.

A second problem which is relevant is the flow and wall dynamics reconstruction using 4D-flow MRI. This imaging modality is richer than Doppler ultrasound and provides directly a measure of the 3D velocity field in the voxels. This enables the use of direct estimation methods at a reduced computational cost with respect to the traditional variational data assimilation approaches. Yet, the sensitivity of the results to subsampling and noise is still not well understood.

We also propose to address the issues related to uncertainty quantification. Indeed, measurements are corrupted by noise and the parameters as well as the available data of the system are either hidden or not known exactly (see [59]). This uncertainty makes the estimation difficult and has a large impact on the precision of the reconstruction, to be quantified in order to provide a reliable tool.

3.2.2. Inverse problem in electro-cardiography

The objective of the inverse problem in electro-cardiography is to recover information about the cardiac electrical activity from electrical measurements on the body surface (for instance from electrocardiograms). We propose to investigate approaches based on recent methods for the Cauchy problem reported in [42]. Basically, the idea consists in regularizing the discrete inverse problem using stabilized finite element methods, without the need of integrating a priori knowledge of the solution, only regularity on the exact solution is required.

3.2.3. Safety pharmacology

One of the the most important problems in pharmacology is cardio-toxicity (see [58]). The objective is to predict whether or not a molecule alters in a significant way the normal functioning of the cardiac cells. This problem can be formulated as inferring the impact of a drug on the ionic currents of each cell based on the measured electrical signal (e.g., electrograms from Micro-Electrodes Arrays). The proposed approach in collaboration with two industrial partners (NOTOCORD and Ncardia) consists in combining available realistic data with virtual ones obtained by numerical simulations. These two datasets can be used to construct efficient classifiers and regressors using machine learning tools (see [41]) and hence providing a rapid way to estimate the impact of a molecule on the electrical activity. The methodological aspects of this work are addressed in Section 3.3.3 .

3.3. Methodological core

The work described in this section is aimed at investigating fundamental mathematical and numerical problems which arise in the first two research axes.

3.3.1. Mathematical analysis of PDEs

The mathematical analysis of the multi-scale and multi-physics models are a fundamental tool of the simulation chain. Indeed, well-posedness results provide precious insights on the properties of solutions of the systems which can, for instance, guide the design of the numerical methods or help to discriminate between different modeling options.

Fluid-structure interaction. Most of the existing results concern the existence of solutions locally in time or away from contacts. One fundamental problem, related to the modeling and simulation of valve dynamics (see Sections 3.1.1 and 3.3.2), is the question of whether or not the model allows for contact (see [57], [55]). The proposed research activity is aimed at investigating the case of both immersed rigid or elastic structures and explore if the considered model allows for contact and if existence can be proved beyond contact. The question of the choice of the model is crucial and considering different types of fluid (newtonian or non newtonian), structure (smooth or rough, elastic, viscoelastic, poro-elastic), or various interface conditions has an influence on whether the model allows contact or not.

Fluid–structure mixture. The main motivation to study fluid-solid mixtures (i.e., porous media consisting of a skeleton and connecting pores filled with fluid) comes from the modeling of the lung parenchyma and cerebral hemorrhages (see Sections 3.1.1 –3.1.2). The Biot model is the most widely used in the literature for the modeling of poro-elastic effects in the arterial wall. Here, we propose to investigate the recent model proposed by the M3DISIM project-team in [47], which allows for nonlinear constitutive behaviors and viscous effects, both in the fluid and the solid. Among the questions which will be addressed, some of them in collaboration with M3DISIM, we mention the justification of the model (or its linearized version) by means of homogenization techniques and its well-posedness.

Fluid–particle interaction. Mathematical analysis studies on the Navier-Stokes-Vlasov system for fluid-particle interaction in aerosols can be found in [38], [39]. We propose to extend these studies to more realistic models which take into account, for instance, changes in the volume of the particles due to humidity.

3.3.2. Numerical methods for multi-physics problems

In this section we describe the main research directions that we propose to explore as regards the numerical approximation of multi-physics problems.

Fluid-structure interaction. The spatial discretization of fluid-structure interaction (FSI) problems generally depends on the amount of solid displacement within the fluid. Problems featuring moderate interface displacements can be successfully simulated using (moving) fitted meshes with an arbitrary Lagrangian-Eulerian (ALE) description of the fluid. This facilitates, in particular, the accurate discretization of the interface conditions. Nevertheless, for problems involving large structural deflections, with solids that might come into contact or that might break up, the ALE formalism becomes cumbersome. A preferred approach in this case is to combine an Eulerian formalism in the fluid with an unfitted mesh discretization, in which the fluid-structure interface deforms independently of a background fluid mesh. In general, traditional unfitted mesh approaches (such as the immersed boundary and the fictitious domain methods [65], [37], [54], [35]) are known to be inaccurate in space. These difficulties have been recently circumvented by a Nitsche-based cut-FEM methodolgy (see [32], [43]). The superior accuracy properties of cut-FEM approaches comes at a price: these methods demand a much more involved computer implementation and require a specific evaluation of the interface intersections.

As regards the time discretization, significant advances have been achieved over the last decade in the development and the analysis of time-splitting schemes that avoid strong coupling (fully implicit treatment of the interface coupling), without compromising stability and accuracy. In the vast majority these studies, the spatial discretization is based on body fitted fluid meshes and the problem of accuracy remains practically open for the coupling with thick-walled structures (see, e.g., [52]). Within the unfitted mesh framework, splitting schemes which avoid strong coupling are much more rare in the literature.

Computational efficiency is a major bottleneck in the numerical simulation of fluid-structure interaction problems with unfitted meshes. The proposed research activity is aimed at addressing these issues. Another fundamental problem that we propose to face is the case of topology changes in the fluid, due to contact or fracture of immersed solids. This challenging problem (fluid-structure-contact-fracture interaction) has major role in many applications (e.g., heart valves repair or replacement, break-up of drug-loaded micro-capsules) but most of the available studies are still merely illustrative. Indeed, besides the numerical issues discussed above, the stability and the accuracy properties of the numerical approximations in such a singular setting are not known.

Fluid–particle interaction and gas diffusion.

Aerosols can be described through mesoscopic equations of kinetic type, which provide a trade-off between model complexity and accuracy. The strongly coupled fluid-particle system involves the incompressible Navier-Stokes equations and the Vlasov equation. The proposed research activity is aimed at investigating the theoretical stability of time-splitting schemes for this system. We also propose to extend these studies to more complex models that take into account the radius growth of the particles due to humidity, and for which stable, accurate and mass conservative schemes have to be developed.

As regards gas diffusion, the mathematical models are generally highly non-linear (see, e.g., [62], [64], [40]). Numerical difficulties arise from these strong non linearities and we propose to develop numerical schemes able to deal with the stiff geometrical terms and that guarantee mass conservation. Moreover, numerical diffusion must be limited in order to correctly capture the time scales and the cross-diffusion effects.

3.3.3. Statistical learning and mathematical modeling interactions

Machine learning and in general statistical learning methods (currently intensively developed and used, see [33]) build a relationship between the system observations and the predictions of the QoI based on the *a posteriori* knowledge of a large amount of data. When dealing with biomedical applications, the

available observations are signals (think for instance to images or electro-cardiograms, pressure and Doppler measurements). These data are high dimensional and the number of available individuals to set up precise classification/regression tools could be prohibitively large. To overcome this major problem and still try to exploit the advantages of statistical learning approaches, we try to add, to the a posteriori knowledge of the available data an *a priori* knowledge, based on the mathematical modeling of the system. A large number of numerical simulations is performed in order to explore a set of meaningful scenarios, potentially missing in the dataset. This *in silico* database of virtual experiments is added to the real dataset: the number of individuals is increased and, moreover, this larger dataset can be used to compute semi-empirical functions to reduce the dimension of the observed signals.

Several investigations have to be carried out to systematically set up this framework. First, often there is not a single mathematical model describing a physiological phenomenon, but hierarchies of model of different complexity. Every model is characterized by a model error. How can this be accounted for? Moreover, several statistical estimators can be set up and eventually combined together in order to improve the estimations (see [70]). Other issues have an actual impact and has to be investigated: what is the optimal number of *in silico* experiments to be added? What are the most relevant scenarios to be simulated in relation to the statistical learning approach considered in order to obtain reliable results? In order to answer to these questions, discussions and collaborations with statistics and machine learning groups have to be developed.

3.3.4. Tensor approximation and HPC

Tensor methods have a recent significant development because of their pertinence in providing a compact representation of large, high-dimensional data. Their applications range from applied mathematics and numerical analysis to machine learning and computational physics. Several tensor decompositions and methods are currently available (see [56]). Contrary to matrices, for tensors of order higher or equal to three, there does not exist, in general, a best low rank approximation, the problem being ill posed (see [68]). Two main points will be addressed: (i) The tensor construction and the multi-linear algebra operations involved when solving high-dimensional problems are still sequential in most of the cases. The objective is to design efficient parallel methods for tensor construction and computations; (ii) When solving high-dimensional problems, the tensor is not assigned; instead, it is specified through a set of equations and tensor data. Our goal is to devise numerical methods able to (dynamically) adapt the rank and the discretization (possibly even the tensor format) to respect the chosen error criterion. This could, in turn, improve the efficiency and reduce the computational burden.

These sought improvements could make the definition of parsimonious discretizations for kinetic theory and uncertainty quantification problems (see Section 3.2.1) more efficient and suitable for a HPC paradigm. This work will be carried out in collaboration with Olga Mula (Université Paris-Dauphine) and the ALPINES and MATHERIALS project-teams.

DRACULA Project-Team

3. Research Program

3.1. Mixed-effect models and statistical approaches

Most of biological and medical data our team has to deal with consist in time series of experimental measurements (cell counts, gene expression level, etc.). The intrinsic variability of any biological system complicates its confrontation to models. The trivial use of means, eliminating the data variance, is but a second-best solution. Furthermore, the amount of data that can be experimentally generated often limits the use of classical mathematical approaches because model's identifiability or parameter identifiability cannot be obtained. In order to overcome this issue and to efficiently take advantage of existing and available data, we plan to use mixed effect models for various applications (for instance: leukemia treatment modeling, immune response modeling). Such models were initially developed to account for individual behaviors within a population by characterizing distributions of parameter values instead of a unique parameter value. We plan to use those approaches both within that frame (for example, taking into account longitudinal studies on different patients, or different mice) but also to extend its validity in a different context: we will consider different ex vivo experiments as being "different individuals": this will allow us to make the most of the experience-to-experience variations.

Such approaches need expertise in statistics to be correctly implemented, and we will rely on the presence of Céline Vial in the team to do so. Céline Vial is an expert in applied statistics and her experience already motivated the use of better statistical methods in various research themes. The increasing use of single cell technologies in biology make such approaches necessary and it is going to be critical for the project to acquire such skills.

3.2. Development of a simulation platform

We have put some effort in developing the *SiMuScale* platform, a software coded in C_{++} dedicated to exploring multiscale population models, since 2014. In order to answer the challenges of multi-scale modeling it is necessary to possess an all-purpose, fast and flexible modeling tool, and *SiMuScale* is the choice we made. Since it is based on a core containing the simulator, and on plug-ins that contain the biological specifications of each cell, this software will make it easier for members of the team – and potentially other modelers – to focus on the model and to capitalize on existing models, which all share the same framework and are compatible with each other. Within the next four years, *SiMuScale* should be widely accessible and daily used in the team for multi-scale modeling. It will be developed into a real-case context, the modeling of the hematopoietic stem cell niche, in collaboration with clinicians (Eric Solary, INSERM) and physicists (Bertrand Laforge, UPMC).

3.3. Mathematical and computational modeling

Multi-scale modeling of hematopoiesis is one of the key points of the project that has started in the early stage of the Dracula team. Investigated by the team members, it took many years of close discussion with biologists to get the best understanding of the key role played by the most important molecules, hormones, kinase cascade, cell communication up to the latest knowledge. An approach that we used is based on hybrid discrete-continuous models, where cells are considered as individual objects, intracellular regulatory networks are described with ordinary differential equations, extracellular concentrations with diffusion or diffusion-convection equations (see Figure 1). These modeling tools require the expertise of all team members to get the most qualitative satisfactory model. The obtained models will be applied particularly to describe normal and pathological hematopoiesis as well as immune response.

3.4. From hybrid dynamics to continuum mechanics

Hybrid discrete-continuous methods are well adapted to describe biological cells. However, they are not appropriate for the qualitative investigation of the corresponding phenomena. Therefore, hybrid model approach should be combined with continuous models. If we consider cell populations as a continuous medium, then cell concentrations can be described by reaction-diffusion systems of equations with convective terms. The diffusion terms correspond to a random cell motion and the reaction terms to cell proliferation, differentiation and death. We will continue our studies of stability, nonlinear dynamics and pattern formation. Theoretical investigations of reaction-diffusion models will be accompanied by numerical simulations and will be applied to study cell population dynamic.

3.5. Structured partial differential equations

Hyperbolic problems are also of importance when describing cell population dynamics. They are structured transport partial differential equations, in which the structure is a characteristic of the considered population, for instance age, size, maturity, etc. In the scope of multi-scale modeling, protein concentrations as structure variables can precisely indicate the nature of cellular events cells undergo (differentiation, apoptosis), by allowing a representation of cell populations in a multi-dimensional space. Several questions are still open in the study of this problem, yet we will continue our analysis of these equations by focusing in particular on the asymptotic behavior of the system (stability, oscillations) and numerical simulations.

3.6. Delay differential equations

The use of age structure in PDE often leads to a reduction (by integration over the age variable) to delay differential equations. Delay differential equations are particularly useful for situations where the processes are controlled through feedback loops acting after a certain time. For example, in the evolution of cell populations the transmission of control signals can be related to some processes as division, differentiation, maturation, apoptosis, etc. Delay differential equations offer good tools to study the behavior of the systems. Our main investigation will be the effect of perturbations of the parameters, as cell cycle duration, apoptosis, differentiation, self-renewal, etc., on the behavior of the system, in relation for instance with some pathological situations. The mathematical analysis of delay differential equations is often complicated and needs the development of new criteria to be performed.

3.7. Multi-scale modeling of the immune response

The main objective of this part is to develop models that make it possible to investigate the dynamics of the adaptive CD8 T cell immune response, and in particular to focus on the consequences of early molecular events on the cellular dynamics few days or weeks later: this would help developing predictive tools of the immune response in order to facilitate vaccine development and reduce costs. This work requires a close and intensive collaboration with immunologist partners.

We recently published a model of the CD8 T cell immune response characterizing differentiation stages, identified by biomarkers, able to predict the quantity of memory cells from early measurements ([40]). In parallel, we improved our multiscale model of the CD8 T cell immune response, by implementing a full differentiation scheme, from naïve to memory cells, based on a limited set of genes and transcription factors.

Our first task will be to infer an appropriate gene regulatory network (GRN) using single cell data analysis (generate transcriptomics data of the CD8 T cell response to diverse pathogens), the previous biomarkers we identified and associated to differentiation stages, as well as piecewise-deterministic Markov processes (Ulysse Herbach's PhD thesis, ongoing).

Our second task will be to update our multiscale model by first implementing the new differentiation scheme we identified ([40]), and second by embedding CD8 T cells with the GRN obtained in our first task (see above). This will lead to a multi-scale model incorporating description of the CD8 T cell immune response both at the molecular and the cellular levels (Simon Girel's PhD thesis, ongoing).

In order to further develop our multiscale model, we will consider an agent-based approach for the description of the cellular dynamics. Yet, such models, coupled to continuous models describing GRN dynamics, are computationally expensive, so we will focus on alternative strategies, in particular on descriptions of the cellular dynamics through both continuous and discrete models, efficiently coupled. Using discrete models for low cell numbers and continuous (partial differential equations) models for large cell numbers, with appropriate coupling strategies, can lead to faster numerical simulations, and consequently can allow performing intense parameter estimation procedures that are necessary to validate models by confronting them to experimental data, both at the molecular and cellular scales.

The final objective will be to capture CD8 T cell responses in different immunization contexts (different pathogens, tumor) and to predict cellular outcomes from molecular events.

3.8. Dynamical network inference from single-cell data

Up to now, all of our multiscale models have incorporated a dynamical molecular network that was build "by hand" after a thorough review of the literature. It would be highly valuable to infer it directly from gene expression data. However, this remains very challenging from a methodological point of view. We started exploring an original solution for such inference by using the information contained within gene expression distributions. Such distributions can be acquired through novel techniques where gene expression levels are quantified at the single cell level. We propose to view the inference problem as a fitting procedure for a mechanistic gene network model that is inherently stochastic and takes not only protein, but also mRNA levels into account. This approach led to very encouraging results [41] and we will actively pursue in that direction, especially in the light of the foreseeable explosion of single cell data.

3.9. Leukemia modeling

Imatinib and other tyrosine kinase inhibitors (TKIs) have marked a revolution in the treatment of Chronic Myelogenous Leukemia (CML). Yet, most patients are not cured, and must take their treatment for life. Deeper mechanistic understanding could improve TKI combination therapies to better control the residual leukemic cell population. In a collaboration with the Hospital Lyon Sud and the University of Maryland, we have developed mathematical models that integrate CML and an autologous immune response ([37], [38] and [39]). These studies have lent theoretical support to the idea that the immune system plays a rôle in maintaining remission over long periods. Our mathematical model predicts that upon treatment discontinuation, the immune system can control the disease and prevent a relapse. There is however a possibility for relapse via a sneak-though mechanism [37]. Research in the next four years will focus in the Phase III PETALS trial. In the PETALS trial (https://clinicaltrials.gov/ct2/show/NCT02201459), the second generation TKI Nilotinib is combined with Peg-IFN, an interferon that is thought to enhance the immune response. We plan to: 1) Adapt the model to take into account the early dynamics (first three months). 2) Use a mixed-effect approach to analyse the effect of the combination, and find population and individual parameters related to treatment efficacy and immune system response. 3) Optimise long-term treatment strategies to reduce or cease treatment and make personalised predictions based on mixed-effect parameters, to minimise the long-term probability of relapse.

DYLISS Project-Team

3. Research Program

3.1. Computer science – symbolic artificial intelligence

We develop methods that use an explicit representation of the relationships between heterogeneous data and knowledge in order to construct a space of hypotheses. Therefore, our objectives in computer science is mainly to develop accurate representations (oriented graphs, Boolean networks, automata, or expressive grammars) to iteratively capture the complexity of a biological system.

Integrating data with querying languages: Semantic web for life sciences The first level of complexity in the data integration process consists in confronting heterogeneous datasets. Both the size and the heretogeneity of life science data make their integration and analysis by domain experts impractical and prone to the streetlight effect (they will pick up the models that best match what they know or what they would like to discover). Our first objective involves the formalization and management of knowledge, that is, the explicitation of relations occurring in structured data. In this setting, our main goal is to facilitate and optimize the integration of Semantic Web resources with local users data by relying on the implicit data scheme contained in biological data and Semantic Web resources.

Reasoning over structured data with constraint-based logical paradigms Another level of complexity in life science integration is that very few paradigms exist to model the behavior of a complex biological system. This leads biologists to perform and formulate hypotheses in order to interpret their data. Our strategy is to interpret such hypotheses as combinatorial optimization problems allowing to reduce the family of models compatible with data. To that goal, we collaborate with Potsdam University in order to use and challenge the most recent developments of Answer Set Programming (ASP) [58], a logical paradigm for solving constraint satisfiability and combinatorial optimization issues. Our goal is therefore to provide scalable and expressive formal models of queries on biological networks with the focus of integrating dynamical information as explicit logical constraints in the modeling process.

Characterizing biological sequences with formal syntactic models Our last goal is to identify and characterize the function of expressed genes in non-model species, such as enzymes and isoforms functions in biological networks or specific functional features of metagenomic samples. These are insufficiently precise because of the divergence of biological sequences, the complexity of molecular structures and biological processes, and the weak signals characterizing these elements. Our goal is therefore to develop accurate formal syntactic models (automata, grammars, abstract gene models) enabling us to represent sequence conservation, sets of short and degenerated patterns and crossing or distant dependencies. This requires both to determine classes of formal syntactic models allowing to handle biological complexity, and to automatically characterize the functional potential embodied in biological sequences with these models.

3.2. Scalable methods to query data heterogenity

Confronted to large and complex data sets (raw data are associated with graphs depicting explicit or implicit links and correlations) almost all scientific fields have been impacted by the *big data issue*, especially genomics and astronomy [67]. In our opinion, life sciences cumulates several features that are very specific and prevent the direct application of big data strategies that proved successful in other domains such as experimental physics: the existence of **several scales of granularity** from microscopic to macroscopic and the associated issue of dependency propagation, datasets **incompleteness and uncertainty** including highly **heterogeneous** responses to a perturbation from one sample to another, and highly fragmented sources of information that **lacks interoperability** [57]. To explore this research field, we use techniques from symbolic data mining (Semantic Web technologies, symbolic clustering, constraint satisfaction and grammatical modelling) to take into account those life science features in the analysis of biological data.

3.2.1. Research topics

Facilitating data integration and querying The quantity and inner complexity of life science data require semantically-rich analysis methods. A major challenge is then to combine data (from local project as well as from reference databases) and symbolic knowledge seamlessly. Semantic Web technologies (RDF for annotating data, OWL for representing symbolic knowledge, and SPARQL for querying) provide a relevant framework, as demonstrated by the success of Linked (Open) Data [44]. However, life science end users (1) find it difficult to learn the languages for representing and querying Semantic Web data, and consequently (2) miss the possibility they had to interact with their tabulated data (even when doing so was exceedingly slow and tedious). Our first objective in this axis is to develop accurate abstractions of datasets or knowledge repositories to facilitate their exploration with RDF-based technologies.

Scalability of semantic web queries. A bottleneck in data querying is given by the performance of federated SPARQL queries, which must be improved by several orders of magnitude to allow current massive data to be analyzed. In this direction, our research program focuses on the combination of *linked data fragments* [68], query properties and dataset structure for decomposing federated SPARQL queries.

Building and compressing static maps of interacting compounds A final approach to handle heterogeneity is to gather multi-scale data knowledge into functional static map of biological models that can be analyzed and/or compressed. This requires to linking genomics, metabolomics, expression data and protein measurement of several phenotypes into unified frameworks. In this direction, our main goal is to develop families of constraints, inspired by symbolic dynamical systems, to link datasets together. We currently focus on health (personalized medicine) and environmental (role of non-coding regulations, graph compression) datasets.

3.2.2. Associated software tools

AskOmics platform *AskOmics* is an integration and interrogation software for linked biological data based on semantic web technologies [url]. *AskOmics* aims at bridging the gap between end user data and the Linked (Open) Data cloud (LOD cloud). It allows heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) to be loaded into a Triple Store system using a user-friendly web interface. It helps end users to (1) take advantage of the information readily available in the LOD cloud for analyzing there own data and (2) contribute back to the linked data by representing their data and the associated metadata in the proper format as well as by linking them to other resources. An originality is the graphical interface that allows any dataset to be integrated in a local RDF datawarehouse and SPARQL query to be built transparently and iteratively by a non-expert user.

FinGoc-tools *The FinGoc tools* allow filtering interaction networks with graph-based optimization criteria in order to elucidate the main regulators of an observed phenotype. The main added-value of these tools is to make explicit the criteria used to highlight the role of the main regulators.(1) The KeyRegulatorFinder package searches key regulators of lists of molecules (like metabolites, enzymes or genes) by taking advantage of knowledge databases in cell metabolism and signaling [package]. (2) The PowerGrasp python package implements graph compression methods oriented toward visualization, and based on power graph analysis [package]. (3) The iggy package enables the repairing of an interaction graph with respect to expression data. [Python package]

3.3. Metabolism: from enzyme sequences to systems ecology

Our researches in bioinformatics in relation with metabolic processes are driven by the understanding of non-model (eukaryote) species. Their metabolism have acquired specific features that we wish to identify with computational methods. To that goal, we combine sequence analysis with metabolic network analysis, with the final goal to understand better the metabolism of communities of organisms.

3.3.1. Research topics

Genomic level: characterizing enzymatic functions of protein sequences Precise characterization of functional proteins, such as enzymes or transporters, is a key to better understand and predict the actors involved in a metabolic process. In order to improve the precision of functional annotations, we develop machine learning approaches taking a sample of functional sequences as input to infer a grammar representing their key syntactical characteristics, including dependencies between residues. Our first goal is to enable an automatic semi-supervised refinement of enzymes classification [6] by combining the Protomata-Learner [50] framework - which captures local dependencies - with formal concept analysis. More challenging, we are exploring the learn of grammars representing long-distance dependencies such as those exhibited by contacts of amino-acids that are far in the sequence but close in the 3D protein folding.

System level: enriching and comparing metabolic networks for non-model organisms Non-model organisms are associated with often incomplete and poorly annotated sequences, leading to draft networks of their metabolism which largely suffer from incompleteness. In former studies, the team has developed several methods to improve the quality of eukaryotes metabolic networks, by solving several variants of the so-called *Metabolic Network gap-filling problem* with logical programming approaches [10], [9]. The main drawback of these approaches is that they cannot scale to the reconstruction and comparison of families of metabolic networks. Our main objective is therefore to develop new tools for the comparison of species strains at the metabolic level.

Consortium level: exploring the diversity of community consortia A new emerging field is system ecology, which aims at building predictive models of species interactions within an ecosystem for deciphering cooperative and competitive relationships between species [56]. This field raises two new issues (1) uncertainty on the species present in the ecosystem and (2) uncertainty about the global objective governing an ecosystem. To address these challenges, our first research focus is the inference of metabolic exchanges and relationships for transporter identification, based on our expertise in metabolic network gap-filling. A second very challenging focus is the prediction of transporters families by obtaining refined characterization of transporters, which are quite unexplored apart from specific databases [65].

3.3.2. Associated software tools

Protomata[url] is a machine learning suite for the inference of automata characterizing (functional) families of proteins at the sequence level. It provides programs to build a new kind of sequences alignments (said partial and local), learn automata and search for new family members in sequence databases. By enabling to model dependencies between positions, automata are more expressive than classical tools (PSSMs, Profile HMMs, or Prosite Patterns) and are well suited to predict new family members with a high specificity. This suite is for instance embedded in the cyanolase database [50] to automate its updade and was used for refining the classification of HAD enzymes [6].

AuReMe workspace is designed for tractable reconstruction of metabolic networks [url]. The toolbox allows for the Automatic Reconstruction of Metabolic networks based on the combination of multiple heterogeneous data and knowledge sources [1]. The main added-values are the inclusion of graph-based tools relevant for the study of non-classical organisms (Meneco and Menetools packages), the possibility to trace the reconstruction and curation procedures (Padmet and Padmet-utils packages), and the exploration of reconstructed metabolic networks with wikis (wiki-export package, see: [url]. It also generated outputs to explore resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae [62], extremophile bacteria [52] and communities of organisms [4].

Mpwt is a Python package for running Pathway Tools [url] on multiple genomes using multiprocessing. Pathway Tools is a comprehensive systems biology software system that is associated with the BioCyc database collection [url]. Pathway Tools is very used for reconstructing metabolic networks.

Metage2metabo is a Python tool to perform graph-based metabolic analysis starting from annotated genomes (reference genomes or metagenome-assembled genomes). It uses Mpwt to reconstruct metabolic networks for a large number of genomes. The obtained metabolic networks are then analyzed individually and collectively in order to get the added value of metabolic cooperation in microbiota over individual metabolism and to identify and screen interesting organisms among all.

3.4. Regulation and signaling: detecting complex and discriminant signatures of phenotypes

On the contrary to metabolic networks, regulatory and signaling processes in biological systems involves agents interacting at different granularity levels (from genes, non-coding RNAs to protein complexes) and different time-scales. Our focus is on the reconstruction of large-scale networks involving multiple scales processes, from which controllers can be extracted with symbolic dynamical systems methods. A particular attention is paid to the characterization of products of genes (such as isoform) and of perturbations to identify discriminant signature of pathologies.

3.4.1. Research topics

Genomic level: characterizing gene structure with grammatical languages and conservation information The subject here is to accurately represent gene structure, including intron/exon structure, for predicting the products of genes, such as isoform transcripts, and comparing the expression potential of a eukaryotic gene according to its context (e.g. tissue) or according to the species. Our approach consists in designing grammatical and comparative-genomics based models for gene structures able to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [46]. Accurate gene models are defined by identifying general constraints shaping gene families and their structures conserved over evolution. Syntactic elements controlling gene expression (transcription factor binding sites controlling transcription; enhancers and silencers controlling splicing events...), i.e. short, degenerated and overlapping functional sequences, are modeled by relying on the high capability of SVG grammars to deal with structure and ambiguity [66].

System level: extracting causal signatures of complex phenotypes with systems biology frameworks The main challenge we address is to set up a generic formalism to model inter-layer interactions in large-scale biological networks. To that goal, we have developed several types of abstractions: multi-experiments framework to learn and control signaling networks [11], multi-layer reactions in interaction graphs [47], and multi-layer information in large-scale Petri nets [43]. Our main issues are to scale these approaches to standardized large-scale repositories by relying on the interoperable Linked Open Data (LOD) resources and to enrich them with ad-hoc regulations extracted from sequence-based analysis. This will allow us to characterize changes in system attractors induced by mutations and how they may be included in pathology signatures.

3.4.2. Associated software tools

Logol software is designed for complex pattern modelling and matching [url]. It is a swiss-army-knife for pattern matching on DNA/RNA/Protein sequences, based on expressive patterns which consist in a complex combination of motifs (such as degenerated strings) and structures (such as imperfect stem-loop ou repeats) [2]. *Logol* key features are the possibilities (i) to divide a pattern description into several sub-patterns, (ii) to model long range dependencies, and (iii) to enable the use of ambiguous models or to permit the inclusion of negative conditions in a pattern definition. Therefore, *Logol* encompasses most of the features of specialized tools (Vmatch, Patmatch, Cutadapt, HMM) and enables interplays between several classes of patterns (motifs and structures), including stem-loop identification in CRISPR.

Caspo software Cell ASP Optimizer (*Caspo*) constitutes a pipeline for automated reasoning on logical signaling networks (learning, classifying, designing experimental perturbations, identifying controllers, take time-series into account) [url]. The software handles inherent experimental noise by enumerating all different logical networks which are compatible with a set of experimental observations [11]. The main advantage is that it enables a complete study of logical network without requiring any linear constraint programs.

Cadbiom package aims at building and analyzing the asynchronous dynamics of enriched logical networks [url] It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [43]. For instance, it was designed to allow controler of phenotypes in large-scale knowledge databases (PID) to be curated and analyzed [5].

EMPENN Project-Team

3. Research Program

3.1. Scientific Foundations

The scientific foundations of our team concern the design and development of new computational solutions for biological images, signals and measurements. Our objective is to develop a better understanding of the normal and pathological brain, at different scales.

This includes imaging brain pathologies in order to better understand pathological behavior from the organ level to the cellular level, and even to the molecular level (using molecule (e.g. through PET-MR imaging), as well as modeling with specific ligands/nanocarriers), and the modelling of normal and pathological large groups of individuals (cohorts) from image descriptors. It also includes the challenge of the discovery of episodic findings (i.e. rare events in large volumes of images and data), data mining and knowledge discovery from image descriptors, the validation and certification of new drugs from imaging features, and, more generally, the integration of neuroimaging into neuroinformatics through the promotion and support of virtual organizations of biomedical actors by means of e-health technologies.

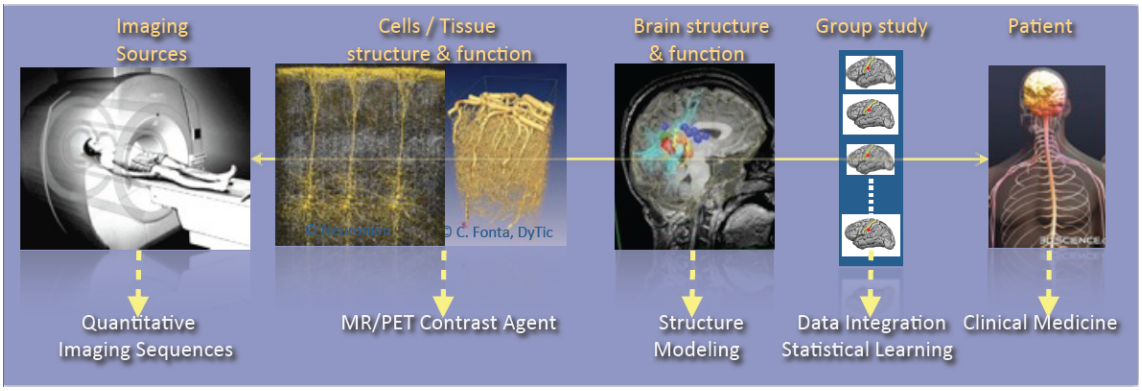


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales: from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions, to the population level for the modelling of group patterns and the learning of group or individual imaging markers.

As shown in Fig. 1 , the research activities of the Empenn team closely link observations and models through the integration of clinical and multiscale data, and phenotypes (cellular, and later molecular, with structural or connectivity patterns in the first stage). Our ambition is to build personalized models of central nervous system organs and pathologies, and to compare these models with clinical research studies in order to establish a quantitative diagnosis, prevent the progression of diseases and provide new digital recovery strategies, while combining all these research areas with clinical validation. This approach is developed within a translational framework, where the data integration process to build the models is informed by specific clinical studies, and where the models are assessed regarding prospective clinical trials for diagnosis and therapy planning. All of these research activities will be conducted in close collaboration with the Neurinfo platform, which benefited in 2018 from a new high-end 3T MRI system dedicated to research (3T Prisma™ system from Siemens), and through the development in the coming years of multimodal hybrid imaging (from the currently available EEG-MRI, to EEG-NIRS and PET-MRI in the future).

In this context, some of our major developments and newly arising issues and challenges will include:

- The generation of new descriptors to study brain structure and function (e.g. the combination of variations in brain perfusion with and without a contrast agent; changes in brain structure in relation to normal, pathological, functional or connectivity patterns; or the modeling of brain state during cognitive stimulation using neurofeedback).
- The integration of additional spatiotemporal and hybrid imaging sequences covering a larger range of observations, from the molecular level to the organ level, via the cellular level (arterial spin labeling, diffusion MRI, MR relaxometry, MR fingerprinting, MR cell labeling imaging, MR-PET molecular imaging, EEG-MRI functional imaging, EEG-NIRS-MRI, etc.).
- The creation of computational models through the data fusion of molecular, cellular (i.e. through dedicated ligands or nanocarriers), structural and functional image descriptors from group studies of normal and/or pathological subjects.
- The evaluation of these models in relation to acute pathologies, especially for the study of degenerative, psychiatric, traumatic or developmental brain diseases (primarily multiple sclerosis, stroke, traumatic brain injury (TBI) and depression, but applicable with a potential additional impact to epilepsy, Parkinson's disease, dementia, Posttraumatic stress disorder, etc.) within a translational framework.

In terms of new major methodological challenges, we will address the development of models and algorithms to reconstruct, analyze and transform the images, and to manage the mass of data to store, distribute and "semanticize" (i.e. provide a logical division of the model's components according to their meaning). As such, we expect to make methodological contributions in the fields of model inference; statistical analysis and modeling; the application of sparse representation (compressed sensing and dictionary learning) and machine learning (supervised/unsupervised classification and discrete model learning); data fusion (multimodal integration, registration, patch analysis, etc.); high-dimensional optimization; data integration; and brain-computer interfaces. As a team at the frontier between the digital sciences and clinical research in neuroscience, we do not claim to provide theoretical breakthroughs in these domains but rather to provide significant advances in using these algorithms through to the advanced applications we intend to address. In addition, we believe that by providing these significant advances using this set of algorithms, we will also contribute to exhibiting new theoretical problems that will fuel the domains of theoretical computer sciences and applied mathematics.

In summary, we expect to address the following major challenges:

- Developing new information processing methods able to detect imaging biomarkers in the context of mental, neurological, and substance use disorders.
- Providing new computational solutions for our target applications, allowing a more appropriate representation of data for image analysis and the detection of biomarkers specific to a form or grade of pathology, or specific to a population of subjects.
- Providing, for our target applications, new patient-adapted connectivity atlases for the study and characterization of diseases from quantitative MRI.
- Providing, for our target applications, new analytical models of dynamic regional perfusion, and deriving indices of dynamic brain local perfusion from normal and pathological populations.
- Investigating whether the theragnostics paradigm of rehabilitation from hybrid neurofeedback can be effective in some behavioral and disability pathologies.

These major advances will be primarily developed and validated in the context of several priority applications in which we expect to play a leading role: multiple sclerosis, stroke rehabilitation, and the study and treatment of depression.

EPIONE Project-Team

3. Research Program

3.1. Introduction

Our research objectives are organized along 5 scientific axes:

1. Biomedical Image Analysis & Machine Learning
2. Imaging & Phenomics, Biostatistics
3. Computational Anatomy, Geometric Statistics
4. Computational Physiology & Image-Guided Therapy
5. Computational Cardiology & Image-Based Cardiac Interventions

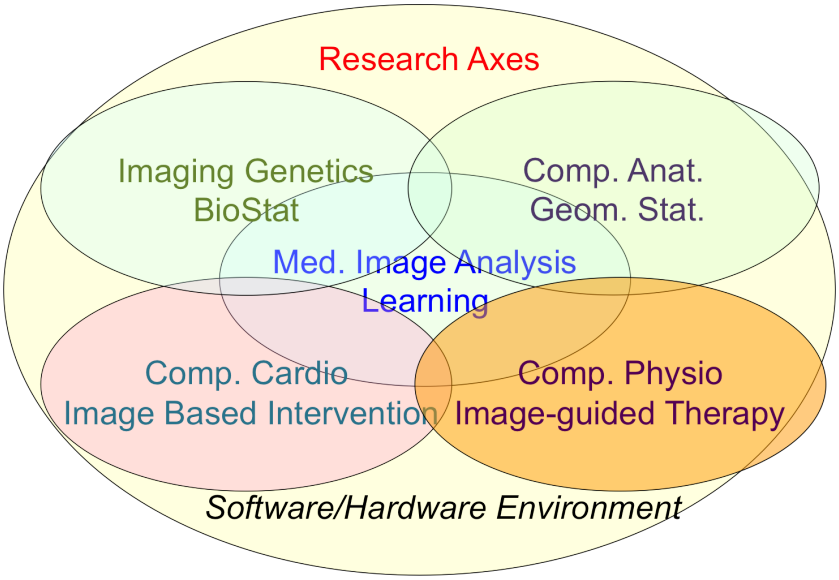


Figure 3. Epione's five main research axes

For each scientific axis, we introduce the context and the long term vision of our research.

3.2. Biomedical Image Analysis & Machine Learning

The long-term objective of biomedical image analysis is to extract, from biomedical images, pertinent information for the construction of the e-patient and for the development of e-medicine. This relates to the development of advanced segmentation and registration of images, the extraction of image biomarkers of pathologies, the detection and classification of image abnormalities, the construction of temporal models of motion or evolution from time-series of images, etc.

A good illustration of the current state of the art and of the remaining challenges can be found in these recent publications which address for instance the extraction of quantitative biomarkers on static or time varying images, as well as image registration and deformation analysis problems. This also applies to the analysis of microscopic and multi-scale images.

In addition, the growing availability of very large databases of biomedical images, the growing power of computers and the progress of machine learning (ML) approaches have opened up new opportunities for biomedical image analysis.

This is the reason why we decided to revisit a number of biomedical image analysis problems with ML approaches, including segmentation and registration problems, automatic detection of abnormalities, prediction of a missing imaging modality, etc. Not only those ML approaches often outperform the previous state-of-the-art solutions in terms of performances (accuracy of the results, computing times), but they also tend to offer a higher flexibility like the possibility to be transferred from one problem to another one with a similar framework. However, even when successful, ML approaches tend to suffer from a lack of explanatory power, which is particularly annoying for medical applications. We also plan to work on methods that can interpret the results of the ML algorithms that we develop.

- **Revisiting Segmentation problems with Machine Learning:** Through a partnership with Microsoft Research in Cambridge (UK), we are studying new segmentation methods based on deep learning with *weakly annotated* data. In effect, a complete segmentation ground truth is costly to collect in medical image analysis, as it requires the tedious task of contouring regions of interest and their validation by an expert. On the other hand, the label "presence" or "absence" of a lesion for instance (weak annotation) can be obtained at a much lower cost.

  We also plan to explore the application of deep learning methods to the fast segmentation of static or deformable organs. For instance we plan to use deep learning methods for the 3D consistent segmentation of the myocardium tissue of the 2 cardiac ventricles, an important preliminary step to mesh the cardiac muscle for computational anatomy, physiology and cardiology projects.
- **Revisiting Registration problems with Machine Learning:** We are studying, through a partnership with Siemens (Princeton), the possibility to apply robust non-rigid registration through agent-based action learning. We propose a decision process where the objective simplifies to iteratively finding the strategically next best step. An artificial agent is driven to solve the task of non-rigid registration through exploring the parametric space of a statistical deformation model built from training data. Since it is difficult to extract trustworthy ground-truth deformation fields we propose a training scheme with synthetically deformed cases and few real inter-subject cases.
- **Prediction of an imaging modality from other imaging modalities with machine learning:** Through a partnership with the Brain and Stem Institute in Paris, we plan to develop deep learning approaches to quantify some brain alterations currently measured by an invasive nuclear medicine imaging modality (PET imaging with specific tracers), directly from a multi-sequence acquisition of a non-invasive imaging modality (MRI). This requires innovative approaches taking into account the relatively small size of the ground truth database (patients having undergone both PET and MR Image acquisitions) and exploiting the a priori knowledge on the brain anatomy. We believe that this approach could apply to other image prediction problems in the longer term.
- **Prediction of cardiac pathologies with machine learning and image simulation:** Following the important work on cardiac image simulation done during the ERC project MedYMA, we are currently able to simulate time-series of images of various cardiac pathologies for which we can vary the parameters of a generative electro-mechanical model. We plan to develop new deep learning methods exploiting both the *shape* and *motion* phenotypes present in the time-series of images to detect and characterize a number of cardiac pathologies, including subtle assynchronies, local ischemia or infarcts.
- **Measuring Brain, Cognition, Behaviour:** We developed a collaborative project MNC3 which is supported by the excellence initiative Idex UCA^{Jedi}. This project gathers partners from Inria, Nice Hospitals (physicians), Nice University, and IPMC (biologists). The goal is to provide a joint analysis of heterogenous data collected on patients with neurological and psychiatric diseases. Those data include medical imaging (mainly MRI), activity (measured by connected wrists or video or microphones), biology/genomics, and clinical information. We want to show the increase in the statistical power of a joint analysis of the data to classify a pathology and to quantify its evolution.

In addition to these mid-term goals, we have applied to two important projects with local clinicians. A project on "Lung cancer", headed by anatomopathologist P. Hofman, to better exploit the joint information coming from imaging and circulating tumoral cells (in collaboration with Median Tech company); and a project "Cluster headache", headed by neurosurgeon D. Fontaine, to better integrate and exploit information coming from imaging, genetics and clinic (in collaboration with Inria Team Athena).

3.3. Imaging & Phenomics, Biostatistics

The human phenotype is associated with a multitude of heterogeneous biomarkers quantified by imaging, clinical and biological measurements, reflecting the biological and patho-physiological processes governing the human body, and essentially linked to the underlying individual genotype.In order to deepen our understanding of these complex relationships and better identify pathological traits in individuals and clinical groups, a long-term objective of e-medicine is therefore to develop the tools for the joint analysis of this heterogeneous information, termed *Phenomics*, within the unified modeling setting of the e-patient.

Ongoing research efforts aim at investigating optimal approaches at the crossroad between biomedical imaging and bioinformatics to exploit this diverse information. This is an exciting and promising research avenue, fostered by the recent availability of large amounts of data from joint imaging and biological studies (such as the UK biobank [0], ENIGMA [0], ADNI [0],...). However, we currently face important methodological challenges, which limit the ability in detecting and understanding meaningful associations between phenotype and biological information.

To date the most common approach to the analysis of the joint variation between the structure and function of organs represented in medical images, and the classical -omics modalities from biology, such as genomics or lipidomics, is essentially based on the massive univariate statistical testing of single candidate features out of the many available. This is for example the case of genome-wide association studies (GWAS) aimed at identifying statistically significant effects in pools consisting of up to millions of genetics variants. Such approaches have known limitations such as multiple comparison problems, leading to underpowered discoveries of significant associations, and usually explain a rather limited amount of data variance. Although more sophisticated machine learning approaches have been proposed, the reliability and generalization of multivariate methods is currently hampered by the low sample size relatively to the usually large dimension of the parameters space.

To address these issues this research axis investigates novel methods for the integration of this heterogeneous information within a parsimonious and unified multivariate modeling framework. The cornerstone of the project consists in achieving an optimal trade-off between modeling flexibility and ability to generalize on unseen data by developing statistical learning methods informed by prior information, either inspired by "mechanistic" biological processes, or accounting for specific signal properties (such as the structured information from spatio-temporal image time series). Finally, particular attention will be paid to the effective exploitation of the methods in the growing Big Data scenario, either in the meta-analysis context, or for the application in large datasets and biobanks.

- **Modeling associations between imaging, clinical, and biological data.** The essential aspect of this research axis concerns the study of data regularization strategies encoding prior knowledge, for the identification of meaningful associations between biological information and imaging phenotype data. This knowledge can be represented by specific biological mechanisms, such as the complex non-local correlation patterns of the -omics encoded in genes pathways, or by known spatio-temporal relationship of the data (such as time series of biological measurements or images). This axis is based on the interaction with research partners in clinics and biology, such as IPMC (CNRS, France), the Lenval Children's Hospital (France), and University College London (UK). This kind of prior information can be used for defining scalable and parsimonious probabilistic regression models. For example, it can provide relational graphs of data interactions that can be modelled by means of

[0] http://www.ukbiobank.ac.uk/
[0] http://enigma.ini.usc.edu/.
[0] http://adni.loni.usc.edu/

Bayesian priors, or can motivate dimensionality reduction techniques and sparse frameworks to limit the effective size of the parameter space. Concerning the clinical application, an important avenue of research will come from the study of the *reduced* representations of the -omics data currently available in clinics, by focusing on the modeling of the disease variants reported in previous genetic findings. The combination of this kind of data with the information routinely available to clinicians, such as medical images and memory tests, has a great potential for leading to improved diagnostic instruments. The translation of this research into clinical practice is carried out thanks to the ongoing collaboration with primary clinical partners such as the University Hospital of Nice (MNC3 partner, France), the Dementia Research Centre of UCL (UK), and the Geneva University Hospital (CH).

- **Learning from collections of biomedical databases.** The current research scenario is characterised by medium/small scale (typically from 50 to 1000 patients) heterogeneous datasets distributed across centres and countries. The straightforward extension of learning algorithms successfully applied to big data problems is therefore difficult, and specific strategies need to be envisioned in order to optimally exploit the available information. To address this problem, we focus on learning approaches to jointly model clinical data localized in different centres. This is an important issue emerging from recent large-scale multi-centric imaging-genetics studies in which partners can only share model parameters (e.g. regression coefficients between specific genes and imaging features), as represented for example by the ENIGMA imaging-genetics study, led by the collaborators at University of Southern California. This problem requires the development of statistical methods for *federated* model estimation, in order to access data hosted in different clinical institutions by simply transmitting the model parameters, that will be in turn updated by using the local available data. This approach is extended to the definition of stochastic optimization strategies in which model parameters are optimized on local datasets, and then summarized in a meta-analysis context. Finally, this project studies strategies for aggregating the information from heterogeneous datasets, accounting for missing modalities due to different study design and protocols. The developed methodology finds important applications within the context of Big Data, for the development of effective learning strategies for massive datasets in the context of medical imaging (such as with the UK biobank), and beyond.

3.4. Computational Anatomy, Geometric Statistics

Computational anatomy is an emerging discipline at the interface of geometry, statistics and image analysis which aims at developing algorithms to model and analyze the biological shape of tissues and organs. The goal is not only to establish generative models of organ anatomies across diseases, populations, species or ages but also to model the organ development across time (growth or aging) and to estimate their variability and link to other functional, genetic or structural information. Computational anatomy is a key component to support computational physiology and is evidently crucial for building the e-patient and to support e-medicine. Pivotal applications include the spatial normalization of subjects in neuroscience (mapping all the anatomies into a common reference system) and atlas to patient registration to map generic knowledge to patient-specific data. Our objectives will be to develop new efficient algorithmic methods to address the emerging challenges described below and to generate precise specific anatomical model in particular for the brain and the heart, but also other organs and structures (e.g. auditory system, lungs, breasts, etc.).

The objects of computational anatomy are often shapes extracted from images or images of labels (segmentation). The observed organ images can also be modeled using registration as the random diffeomorphic deformation of an unknown template (i.e. an orbit).In these cases as in many other applications, invariance properties lead us to consider that these objects belong to non-linear spaces that have a geometric structure. Thus, the mathematical foundations of computational anatomy rely on statistics on non-linear spaces.

- **Geometric Statistics** aim at studying this abstracted problem at the theoretical level. Our goal is to advance the fundamental knowledge in this area, with potential applications to new areas outside of medical imaging. Several challenges which constitute shorter term objectives in this direction are described below.

- **Large databases and longitudinal evolution:** The emergence of larger databases of anatomical images (ADNI, UK biobank) and the increasing availability of temporal evolution drives the need for efficient and scalable statistical techniques. A key issue is to understand how to construct hierarchical models in a non-linear setting.
- **Non-parametric models of variability:** Despite important successes, anatomical data also tend to exhibit a larger variability than what can be modeled with a standard multivariate unimodal Gaussian model. This raises the need for new statistical models to describe the anatomical variability like Bayesian statistics or sample-based statistical model like multi-atlas and archetypal techniques. A second objective is thus to develop efficient algorithmic methods for encoding the statistical variability into models.
- **Intelligible reduced-order models:** Last but not least, these statistical models should live in low dimensional spaces with parameters that can be interpreted by clinicians. This requires of course dimension reduction and variable selection techniques. In this process, it is also fundamental to align the selected variable to a dictionary of clinically meaningful terms (an ontology), so that the statistical model can not only be used to predict but also to explain.

3.4.1. Geometric Statistics

- **Foundations of statistical estimation on geometric spaces:** Beyond the now classical Riemannian spaces, this axis will develop the foundations of statistical estimation on affine connection spaces (e.g. Lie groups), quotient and stratified metric spaces (e.g. orbifolds and tree spaces). In addition to the curvature, one of the key problem is the introduction of singularities at the boundary of the regular strata (non-smooth and non-convex analysis).
- **Parametric and non-parametric dimension reduction methods in non-linear spaces:** The goal is to extend what is currently done with the Fréchet mean (i.e. a 0-dimensional approximation space) to higher dimensional subspaces and finally to a complete hierarchy of embedded subspaces (flags) that iteratively model the data with more and more precision. The Barycentric Subspace Analysis (BSA) generalization of principal component analysis which was recently proposed in the team will of course be a tool of choice for that. In this process, a key issue is to estimate efficiently not only the model parameters (mean point, subspace, flag) but also their uncertainty. Here, we want to quantify the influence of curvature and singularities on non-asymptotic estimation theory since we always have a finite (and often too limited) number of samples. As the mean is generally not unique in curved spaces, this also leads to consider that the results of estimation procedures should be changed from points to singular distributions. A key challenge in developing such a geometrization of statistics will not only be to unify the theory for the different geometric structures, but also to provide efficient practical algorithms to implement them.
- **Learning the geometry from the data:** Data can be efficiently approximated with locally Euclidean spaces when they are very finely sampled with respect to the curvature (big data setting). In the high dimensional low sample size (small data) setting, we believe that invariance properties are essential to reasonably interpolate and approximate. New apparently antagonistic notions like approximate invariance could be the key to this interaction between geometry and learning.

  Beyond the traditional statistical survey of the anatomical shapes that is developed in computational anatomy above, we intend to explore other application fields exhibiting geometric but non-medical data. For instance, applications can be found in Brain-Computer Interfaces (BCI), tree-spaces in phylogenetics, Quantum Physics, etc.

3.5. Computational Physiology & Image-Guided Therapy

Computational Physiology aims at developing computational models of human organ *functions*, an important component of the e-patient, with applications in e-medicine and more specifically in computer-aided prevention, diagnosis, therapy planning and therapy guidance. The focus of our research is on *descriptive* (allowing

to reproduce available observations), *discriminative* (allowing to separate two populations), and above all *predictive models* which can be personalized from patient data including medical images, biosignals, biological information and other available metadata. A key aspect of this scientific axis is therefore the coupling of biophysical models with patient data which implies that we are mostly considering models with relatively few and identifiable parameters. To this end, *data assimilation* methods aiming at estimating biophysical model parameters in order to reproduce available patient data are preferably developed as they potentially lead to predictive models suitable for therapy planning.

Previous research projects in computational physiology have led us to develop biomechanical models representing quasi-static small or large soft tissue deformations (e.g. liver or breast deformation after surgery), mechanical growth or atrophy models (e.g. simulating brain atrophy related to neurodegenerative diseases), heat transfer models (e.g. simulating radiofrequency ablation of tumors), and tumor growth models (e.g. brain or lung tumor growth).

To improve the data assimilation of biophysical models from patient data, a long term objective of our research will be to develop *joint imaging and biophysical generative models in a probabilistic framework* which simultaneously describe the appearance and function of an organ (or its pathologies) in medical images. Indeed, current approaches for the personalization of biophysical models often proceed in two separate steps. In a first stage, geometric, kinematic or/ functional features are first extracted from medical images. In a second stage, they are used by personalization methods to optimize model parameters in order to match the extracted features. In this process, subtle information present in the image which could be informative for biophysical models is often lost which may lead to limited personalization results. Instead, we propose to develop more integrative approaches where the extraction of image features would be performed jointly with the model parameter fitting. Those imaging and biophysical generative models should lead to a *better understanding* of the content of images, to a *better personalization* of model parameters and also *better estimates of their uncertainty*.

This improved coupling between images and model should *help solving various practical problems* driven by clinical applications. Depending on available resources, datasets, and clinical problems, we wish to develop a new expertise for the simulation of *tissue perfusion* (e.g. to capture the uptake of contrast agent or radioactive tracers), or *blood flow in medium / small vessels* (e.g. to capture the transport of drugs or radioactive materials in interventional radiology).

- **Reduced Computational Biophysical Models**. Clinical constraint and uncertainty estimation inevitably lead to the requirement of relatively fast computation of biophysical models. In addition to hardware acceleration (GPU, multithreading) we will explore various ways to accelerate the computation of models through intrusive (e.g. proper orthogonal decomposition, computation of condensed stiffness matrices in non-linear mechanics) or non intrusive methods (e.g. polynomial chaos expansion, Gaussian processes).
- **Uncertainty estimation of Biophysical Models**. We will pursue our research on this topic by developing Bayesian methods to estimate the posterior probability of model parameters, initial and boundary conditions from image features or image voxels. Such approaches rely on the definition of relevant likelihood terms relating the model state variables to the observable quantities in images. When possible joint imaging and biophysical generative models will be developed to avoid to rely on intermediate image features. Approximate inference of uncertainty will be estimated through Variational Bayes approaches whose accuracy will be evaluated through a comparison with stochastic sampling methods (e.g. MCMC). Through this uncertainty estimation, we also aim at developing a reliable framework to select the most sensitive and discriminative parameters of a given model but also to select the biophysical model best suited to solve a given problem (e.g. prediction of therapy outcome).
- **High Order Finite Element Modeling**. Soft tissue biomechanical models have until now been formulated as linear elastic or hyperelastic materials discretized as linear tetrahedra finite elements. While being very generic, those elements are known to suffer from numerical locking for nearly

incompressible materials and lead to poor estimate of stress field. We will develop efficient implementation and assembly methods using high order tetrahedral (and possibly hexahedral) elements. To maintain the number of nodes relatively low while keeping a good accuracy, we intend to develop elements of adaptive degree (p-refinement) driven by local error indices. Solution for meshing surfaces or volumes with curved high order elements will be developed in collaboration with the Titane and Aromath Inria teams.

- **Clinical Applications**. We plan to develop new applications of therapy planning and therapy guidance through existing or emerging collaborations related to the following problems : breast reconstruction following insertion of breast implants (with Anatoscope), planning of cochlear electrodes implantation (with CHU Nice and Oticon Medical), lung deformation following COPD or pulmonary fibrosis (with CHU Nice), echography based elastometry (with CHU Nice).

3.6. Computational Cardiology & Image-Based Cardiac Interventions

Computational Cardiology has been an active research topic within the Computational Anatomy and Computational Physiology axes of the previous Asclepios project, leading to the development of personalized computational models of the heart designed to help characterizing the cardiac function and predict the effect of some device therapies like cardiac resynchronisation or tissue ablation. This axis of research has now gained a lot of maturity and a critical mass of involved scientists to justify an individualized research axis of the new project Epione, while maintaining many constructive interactions with the 4 other research axes of the project. This will develop all the cardiovascular aspects of the e-patient for cardiac e-medicine.

The new challenges we want to address in computational cardiology are related to the introduction of new levels of modeling and to new clinical and biological applications. They also integrate the presence of new sources of measurements and the potential access to very large multimodal databases of images and measurements at various spatial and temporal scales.

Our goal will be to combine two complementary computational approaches: *machine learning* and *biophysical modelling*. This research axis will leverage on the added value of such a combination. Also we will refine our biophysical modeling by the introduction of a pharmacokinetics/pharmacodynamics (PK/PD) component able to describe the effect of a drug on the cardiac function. This will come in complement to the current geometric, electrical, mechanical and hemodynamic components of our biophysical model of the heart. We will also carefully model the uncertainty in our modeling, and try to provide algorithms fast enough to allow future clinical translation.

- Physics of Ultrasound Images for Probe Design: we will design a digital phantom of the human torso in order to help the design of echocardiographic probes. This will be done in collaboration with GE Healthcare whose excellence centre for cardiac ultrasound probes is located in Sophia Antipolis.
- Cardiac Pharmacodynamics for Drug Personalisation: we will add to our biophysical cardiac model a pharmacodynamics model, coupled with a pharmacokinetics model and a personalisation framework in order to help the adjustment of drug therapy to a given patient. This will be done in collaboration with ExactCure, a start up company specialised on this topic.
- New Imaging Modality Coupling MRI and Electrodes: we will use our fast models in order to regularize the ill-posed inverse problem of cardiac electrocardiography in order to estimate cardiac electrical activity from body surface potentials. This will be done within the ERC Starting Grant ECSTATIC coordinated by Hubert Cochet from the IHU Liryc, Bordeaux.
- Cardiac Imaging during Exercise: a particular aspect of the cardiac function is its constant adaptation to satisfy the needs of the human body. This dynamic aspect provides important information on the cardiac function but is challenging to measure. We will set up exercise protocols with Nice University Hospital and STAPS in order to model and quantify such an adaptation of the cardiac function.
- Sudden Cardiac Death is the cause of important mortality (300 000 per year in Europe, same in US) and it is difficult to identify people at risk. Based on a large multi-centric database of images, we will learn the image features correlated with a high risk of arrhythmia, with the IHU Liryc.

- Personalising models from connected objects: with the Internet of Things and the pletora of sensors available today, the cardiac function can be monitored almost continuously. Such new data open up possibilities for novel methods and tools for diagnosis, prognosis and therapy.

ERABLE Project-Team

3. Research Program

3.1. Two main goals

ERABLE has two main goals, one related to biology and the other to methodology (algorithms, combinatorics, statistics). In relation to biology, the main goal of ERABLE is to contribute, through the use of mathematical models and algorithms, to a better understanding of close and often persistent interactions between "collections of genetically identical or distinct self-replicating cells" which will correspond to organisms/species or to actual cells. The first will cover the case of what has been called symbiosis, meaning when the interaction involves different species, while the second will cover the case of a (cancerous) tumour which may be seen as a collection of cells which suddenly disrupts its interaction with the other (collections of) cells in an organism by starting to grow uncontrollably.

Such interactions are being explored initially at the molecular level. Although we rely as much as possible on already available data, we intend to also continue contributing to the identification and analysis of the main genomic and systemic (regulatory, metabolic, signalling) elements involved or impacted by an interaction, and how they are impacted. We started going to the population and ecological levels by modelling and analysing the way such interactions influence, and are or can be influenced by the ecosystem of which the "collections of cells" are a part. The key steps are:

- identifying the molecular elements based on so-called omics data (genomics, transcriptomics, metabolomics, proteomics, etc.): such elements may be gene/proteins, genetic variations, (DNA/RNA/protein) binding sites, (small and long non coding) RNAs, etc.
- simultaneously inferring and analysing the network that models how these molecular elements are physically and functionally linked together for a given goal, or find themselves associated in a response to some change in the environment;
- modelling and analysing the population and ecological network formed by the "collections of cells in interaction", meaning modelling a network of networks (previously inferred or as already available in the literature).

One important longer term goal of the above is to analyse how the behaviour and dynamics of such a network of networks might be controlled by modifying it, including by subtracting some of its components from the network or by adding new ones.

In relation to methodology, the main goal is to provide those enabling to address our main biological objective as stated above that lead to the best possible interpretation of the results within a given pre-established model and a well defined question. Ideally, given such a model and question, the method is exact and also exhaustive if more than one answer is possible. Three aspects are thus involved here: establishing the model within which questions can and will be put; clearly defining such questions; exactly answering to them or providing some guarantee on the proximity of the answer given to the "correct" one. We intend to continue contributing to these three aspects:

- at the modelling level, by exploring better models that at a same time are richer in terms of the information they contain (as an example, in the case of metabolism, using hypergraphs as models for it instead of graphs) and are susceptible to an easier treatment:
  - these two objectives (rich models that are at the same time easy to treat) might in many cases be contradictory and our intention is then to contribute to a fuller characterisation of the frontiers between the two;
  - even when feasible, the richer models may lack a full formal characterisation (this is for instance the case of hypergraphs) and our intention is then to contribute to such a characterisation;

- at the question level, by providing clear formalisations of those that will be raised by our biological concerns;
- at the answer level:
  - to extend the area of application of exact algorithms by: (i) a better exploration of the combinatorial properties of the models, (ii) the development of more efficient data structures, (iii) a smarter traversal of the space of solutions when more than one solution exists;
  - when exact algorithms are not possible, or when there is uncertainty in the input data to an algorithm, to improve the quality of the results given by a deeper exploration of the links between different algorithmic approaches: combinatorial, randomised, stochastic.

3.2. Different research axes

The goals of the team are biological and methodological, the two being intrinsically linked. Any division into axes along one or the other aspect or a combination of both is thus somewhat artificial. Following the evaluation of the team at the end of 2017, four main axes were identified, with the last one being the more recently added one. This axis is specifically oriented towards health in general, human or animal. The first three axes are: genomics, metabolism and post-transcriptional regulation, and (co)evolution.

Notice that the division itself is based on the biological level (genomic, metabolic/regulatory, evolutionary) or main current Life Science purpose (health) rather than on the mathematical or computational methodology involved. Any choice has its part of arbitrariness. Through the one we made, we wished to emphasise the fact that the area of application of ERABLE is important for us. *It does not mean that the mathematical and computational objectives are not equally important*, but only that those are, most often, motivated by problems coming from or associated to the general Life Science goal. Notice that such arbitrariness also means that some Life Science topics will be artificially split into two different Axes. One example of this is genomics and the main health areas currently addressed that are intrinsically inter-related.

Axis 1: Genomics
Intra and inter-cellular interactions involve molecular elements whose identification is crucial to understand what governs, and also what might enable to control such interactions. For the sake of clarity, the elements may be classified in two main classes, one corresponding to the elements that allow the interactions to happen by moving around or across the cells, and another that are the genomic regions where contact is established. Examples of the first are non coding RNAs, proteins, and mobile genetic elements such as (DNA) transposons, retro-transposons, insertion sequences, etc. Examples of the second are DNA/RNA/protein binding sites and targets. Furthermore, both types (effectors and targets) are subject to variation across individuals of a population, or even within a single (diploid) individual. Identification of these variations is yet another topic that we wish to cover. Variations are understood in the broad sense and cover single nucleotide polymorphisms (SNPs), copy-number variants (CNVs), repeats other than mobile elements, genomic rearrangements (deletions, duplications, insertions, inversions, translocations) and alternative splicings (ASs). All three classes of identification problems (effectors, targets, variations) may be put under the general umbrella of genomic functional annotation.

Axis 2: Metabolism and post-transcriptional regulation
As increasingly more data about the interaction of molecular elements (among which those described above) becomes available, these should then be modelled in a subsequent step in the form of networks. This raises two main classes of problems. The first is to accurately infer such networks. Assuming such a network, integrated or “simple”, has been inferred for a given organism or set of organisms, the second problem is then to develop the appropriate mathematical models and methods to extract further biological information from such networks.

The team has so far concentrated its efforts on two main aspects concerning such interactions: metabolism and post-transcriptional regulation by small RNAs. The more special niche we have been exploring in relation to metabolism concerns the fact that the latter may be seen as an organism's immediate window into its environment. Finely understanding how species communicate through those windows, or what impact they may have on each other through them is thus important when the ultimate goal is to be able to model communities of organisms, for understanding them and possibly, on a longer term, for control. While such communication has been explored in a number of papers, most do so at a too high level or only considered couples of interacting organisms, not larger communities. The idea of investigating consortia, and in the case of synthetic biology, of using them, has thus started being developed in the last decade only, and was motivated by the fact that such consortia may perform more complicated functions than could single populations, as well as be more robust to environmental fluctuations. Another originality of the work that the team has been doing in the last decade has also been to fully explore the combinatorial aspects of the structures used (graphs or directed hypergraphs) and of the associated algorithms. As concerns post-transcriptional regulation, the team has essentially been exploring the idea that small RNAs may have an important role in the dialog between different species.

Axis 3: (Co)Evolution
Understanding how species that live in a close relationship with others may (co)evolve requires understanding for how long symbiotic relationships are maintained or how they change through time. This may have deep implications in some cases also for understanding how to control such relationships, which may be a way of controlling the impact of symbionts on the host, or the impact of the host on the symbionts and on the environment (by acting on its symbiotic partner(s)). These relationships, also called *symbiotic associations*, have however not yet been very widely studied, at least not at a large scale.

One of the problems is getting the data, meaning the trees for hosts and symbionts but even prior to that, determining with which symbionts the present-day hosts are associated (or are "infected" by as may be the term used in some contexts) which is a big enterprise in itself. The other problem is measuring the stability of the association. This has generally been done by concomitantly studying the phylogenies of hosts and symbionts, that is by doing what is called a *cophylogeny* analysis, which itself is often realised by performing what is called a *reconciliation* of two phylogenetic trees (in theory, it could be more than two but this is a problem that has not yet been addressed by the team), one for the symbionts and one for the hosts with which the symbionts are associated. This consists in mapping one of the trees (usually, the symbiont tree) to the other. Cophylogeny inherits all the difficulties of phylogeny, among which the fact that it is not possible to check the result against the "truth" as this is now lost in the past. Cophylogeny however also brings new problems of its own which are to estimate the frequency of the different types of events that could lead to discrepant evolutionary histories, and to estimate the duration of the associations such events may create.

Axis 4: Human, animal and plant health
As indicated above, this is a recent axis in the team and concerns various applications to human and animal health. In some ways, it overlaps with the three previous axes as well as with Axis 5 on the methodological aspects, but since it gained more importance in the past few years, we decided to develop more these particular applications. Most of them started through collaborations with clinicians. Such applications are currently focused on three different topics: (i) Infectiology, (ii) Rare diseases, and (iii) Cancer.

Infectiology is the oldest one. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil that focused on pathogenic bacteria living inside the respiratory tract of swines. Since our participation in the H2020 ITN MicroWine, we started interested in infections affecting plants this time, and more particularly vine plants. Rare Diseases on the other hand started by a collaboration with clinicians from the Centre de Recherche en Neurosciences of Lyon (CNRL) and is focused on the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

The latter collaboration was initiated through a relationship between a member of ERABLE (Alain Viari) and Dr. Gilles Thomas who had been friends since many years. G. Thomas was one of the pioneers of Cancer Genomics in France. After his death in 2014, Alain Viari took the (part time) responsibility of his team at CLB and pursued the main projects he had started.

Within Inria and beyond, the first two applications (Infectiology and Rare Diseases) may be seen as unique because of their specific focus (resp. respiratory tract of swines / vine plants on one hand, and TALS on the other). In the first case, such uniqueness is also related to the fact that the work done involves a strong computational part but also experiments *performed within ERABLE itself*.

FLUMINANCE Project-Team

3. Research Program

3.1. Estimation of fluid characteristic features from images

The measurement of fluid representative features such as vector fields, potential functions or vorticity maps, enables physicists to have better understanding of experimental or geophysical fluid flows. Such measurements date back to one century and more but became an intensive subject of research since the emergence of correlation techniques [56] to track fluid movements in pairs of images of a particles laden fluid or by the way of clouds photometric pattern identification in meteorological images. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, referred to in the literature as optical-flow, is an intensive subject of researches since the 80's and the seminal work of B. Horn and B. Schunk [67]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, p ollutant, light index refraction, fluorescein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally constraints to be included such as kinematic properties or dynamic laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation procedures on the basis of physically grounded data model that describes the relation linking the observed luminance function and some state variables of the observed flow. The Fluminance group has allowed a substantial progress in this direction with the design of dedicated dense estimation techniques to estimate dense fluid motion fields. See [7] for a detailed review. More recently problems related to scale measurement and uncertainty estimation have been investigated [61]. Dynamically consistent and highly robust techniques have been also proposed for the recovery of surface oceanic streams from satellite images [58]. Very recently parameter-free approaches relying on uncertainty concept has been devised [59]. This technique outperforms the state of the art.

3.2. Data assimilation and Tracking of characteristic fluid features

Real flows have an extent of complexity, even in carefully controlled experimental conditions, which prevents any set of sensors from providing enough information to describe them completely. Even with the highest levels of accuracy, space-time coverage and grid refinement, there will always remain at least a lack of resolution and some missing input about the actual boundary conditions. This is obviously true for the complex flows encountered in industrial and natural conditions, but remains also an obstacle even for standard academic flows thoroughly investigated in research conditions.

This unavoidable deficiency of the experimental techniques is nevertheless more and more compensated by numerical simulations. The parallel advances in sensors, acquisition, treatment and computer efficiency allow the mixing of experimental and simulated data produced at compatible scales in space and time. The inclusion of dynamical models as constraints of the data analysis process brings a guaranty of coherency based on fundamental equations known to correctly represent the dynamics of the flow (e.g. Navier Stokes equations) [11]. Conversely, the injection of experimental data into simulations ensures some fitting of the model with reality.

To enable data and models coupling to achieve its potential, some difficulties have to be tackled. It is in particular important to outline the fact that the coupling of dynamical models and image data are far from being straightforward. The first difficulty is related to the space of the physical model. As a matter of fact, physical models describe generally the phenomenon evolution in a 3D Cartesian space whereas images provides generally only 2D tomographic views or projections of the 3D space on the 2D image plane. Furthermore, these views are sometimes incomplete because of partial occlusions and the relations between the model state variables and the image intensity function are otherwise often intricate and only partially known. Besides, the dynamical model and the image data may be related to spatio-temporal scale spaces of very different natures which increases the complexity of an eventual multiscale coupling. As a consequence of these difficulties, it is necessary generally to define simpler dynamical models in order to assimilate image data. This redefinition can be done for instance on an uncertainty analysis basis, through physical considerations or by the way of data based empirical specifications. Such modeling comes to define inexact evolution laws and leads to the handling of stochastic dynamical models. The necessity to make use and define sound approximate models, the dimension of the state variables of interest and the complex relations linking the state variables and the intensity function, together with the potential applications described earlier constitute very stimulating issues for the design of efficient data-model coupling techniques based on image sequences.

On top of the problems mentioned above, the models exploited in assimilation techniques often suffer from some uncertainties on the parameters which define them. Hence, a new emerging field of research focuses on the characterization of the set of achievable solutions as a function of these uncertainties. This sort of characterization indeed turns out to be crucial for the relevant analysis of any simulation outputs or the correct interpretation of operational forecasting schemes. In this context, stochastic modeling play a crucial role to model and process uncertainty evolution along time. As a consequence, stochastic parameterization of flow dynamics has already been present in many contributions of the Fluminance group in the last years and will remain a cornerstone of the new methodologies investigated by the team in the domain of uncertainty characterization.

This wide theme of research problems is a central topic in our research group. As a matter of fact, such a coupling may rely on adequate instantaneous motion descriptors extracted with the help of the techniques studied in the first research axis of the FLUMINANCE group. In the same time, this coupling is also essential with respect to visual flow control studies explored in the third theme. The coupling between a dynamics and data, designated in the literature as a Data Assimilation issue, can be either conducted with optimal control techniques [68], [69] or through stochastic filtering approaches [62], [65]. These two frameworks have their own advantages and deficiencies. We rely indifferently on both approaches.

3.3. Optimization and control of fluid flows with visual servoing

Fluid flow control is a recent and active research domain. A significant part of the work carried out so far in that field has been dedicated to the control of the transition from laminarity to turbulence. Delaying, accelerating or modifying this transition is of great economical interest for industrial applications. For instance, it has been shown that for an aircraft, a drag reduction can be obtained while enhancing the lift, leading consequently to limit fuel consumption. In contrast, in other application domains such as industrial chemistry, turbulence phenomena are encouraged to improve heat exchange, increase the mixing of chemical components and enhance chemical reactions. Similarly, in military and civilians applications where combustion is involved, the control of mixing by means of turbulence handling rouses a great interest, for example to limit infra-red signatures of fighter aircraft.

Flow control can be achieved in two different ways: passive or active control. Passive control provides a permanent action on a system. Most often it consists in optimizing shapes or in choosing suitable surfacing (see for example [60] where longitudinal riblets are used to reduce the drag caused by turbulence). The main problem with such an approach is that the control is, of course, inoperative when the system changes. Conversely, in active control the action is time varying and adapted to the current system's state. This approach requires an external energy to act on the system through actuators enabling a forcing on the flow through for instance blowing and suction actions [72], [64]. A closed-loop problem can be formulated as an optimal control

issue where a control law minimizing an objective cost function (minimization of the drag, minimization of the actuators power, etc.) must be applied to the actuators [57]. Most of the works of the literature indeed comes back to open-loop control approaches [71], [66], [70] or to forcing approaches [63] with control laws acting without any feedback information on the flow actual state. In order for these methods to be operative, the model used to derive the control law must describe as accurately as possible the flow and all the eventual perturbations of the surrounding environment, which is very unlikely in real situations. In addition, as such approaches rely on a perfect model, a high computational costs is usually required. This inescapable pitfall has motivated a strong interest on model reduction. Their key advantage being that they can be specified empirically from the data and represent quite accurately, with only few modes, complex flows' dynamics. This motivates an important research axis in the Fluminance group.

3.4. Numerical models applied to hydrogeology and geophysics

The team is strongly involved in numerical models for hydrogeology and geophysics. There are many scientific challenges in the area of groundwater simulations. This interdisciplinary research is very fruitful with cross-fertilizing subjects.

In geophysics, a main concern is to solve inverse problems in order to fit the measured data with the model. Generally, this amounts to solve a linear or nonlinear least-squares problem.

Models of geophysics are in general coupled and multi-physics. For example, reactive transport couples advection-diffusion with chemistry. Here, the mathematical model is a set of nonlinear Partial Differential Algebraic Equations. At each timestep of an implicit scheme, a large nonlinear system of equations arise. The challenge is to solve efficiently and accurately these large nonlinear systems.

3.5. Numerical algorithms and high performance computing

Linear algebra is at the kernel of most scientific applications, in particular in physical or chemical engineering. The objectives are to analyze the complexity of these different methods, to accelerate convergence of iterative methods, to measure and improve the efficiency on parallel architectures, to define criteria of choice.

GENSCALE Project-Team

3. Research Program

3.1. Axis 1: Data Structures

The aim of this axis is to develop efficient data structures for representing the mass of genomic data generated by the sequencing machines. This research is motivated by the fact that the treatments of large genomes, such as mammalian or plant genomes, or multiple genomes coming from a same sample as in metagenomics, require high computing resources, and more specifically very important memory configuration. The last advances in TGS technologies bring also new challenges to represent or search information based on sequencing data with high error rate.

Part of our research focuses on the de-Bruijn graph structure. This well-known data structure, directly built from raw sequencing data, have many properties matching perfectly well with NGS processing requirements. Here, the question we are interested in is how to provide a low memory footprint implementation of the de-Bruijn graph to process very large NGS datasets, including metagenomic ones [3], [4].

A correlated research direction is the indexing of large sets of objects. A typical, but non exclusive, need is to annotate nodes of the de-Bruijn graph, that is potentially billions of items. Again, very low memory footprint indexing structures are mandatory to manage a very large quantity of objects [7].

3.2. Axis 2: Algorithms

The main goal of the GenScale team is to develop optimized tools dedicated to genomic data processing. Optimization can be seen both in terms of space (low memory footprint) and in terms of time (fast execution time). The first point is mainly related to advanced data structures as presented in the previous section (axis 1). The second point relies on new algorithms and, when possible implementation on parallel structures (axis 3).

We do not have the ambition to cover the vast panel of software related to genomic data processing needs. We particularly focused on the following areas:

- **NGS data Compression** De-Bruijn graphs are de facto a compressed representation of the NGS information from which very efficient and specific compressors can be designed. Furthermore, compressing the data using smart structures may speed up some downstream graph-based analyses since a graph structure is already built [1].
- **Genome assembly** This task remains very complicated, especially for large and complex genomes, such as plant genomes with polyploid and highly repeated structures. We worked both on the generation of contigs [3] and on the scaffolding step [5]. Both NGS and TGS technologies are taken into consideration, either independently or using combined approaches.
- **Detection of variants** This is often the main information one wants to extract from the sequencing data. Variants range from SNPs or short indels to structural variants that are large insertions/deletions and long inversions over the chromosomes. We developed original methods to find variants without any reference genome [10], to detect structural variants using local NGS assembly approaches [9] or TGS processing.
- **Metagenomics** We focused our research on comparative metagenomics by providing methods able to compare hundreds of metagenomic samples together. This is achieved by combining very low memory data structures and efficient implementation and parallelization on large clusters [2].

3.3. Axis 3: Parallelism

This third axis investigates a supplementary way to increase performances and scalability of genomic treatments. There are many levels of parallelism that can be used and/or combined to reduce the execution time of very time-consuming bioinformatics processes. A first level is the parallel nature of today processors that now house several cores. A second level is the grid structure that is present in all bioinformatics centers or in the cloud. This two levels are generally combined: a node of a grid is often a multicore system. Another possibility is to add hardware accelerators to a processor. A GPU board is a good example.

GenScale does not do explicit research on parallelism. It exploits the capacity of computing resources to support parallelism. The problem is addressed in two different directions. The first is an engineering approach that uses existing parallel tools to implement algorithms such as multithreading or MapReduce techniques [4]. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria [2]. This is particularly true for parallel algorithms targeting hardware accelerators.

IBIS Project-Team

3. Research Program

3.1. Analysis of qualitative dynamics of gene regulatory networks

Participants: Hidde de Jong [Correspondent], Michel Page, Delphine Ropers.

The dynamics of gene regulatory networks can be modeled by means of ordinary differential equations (ODEs), describing the rate of synthesis and degradation of the gene products as well as regulatory interactions between gene products and metabolites. In practice, such models are not easy to construct though, as the parameters are often only constrained to within a range spanning several orders of magnitude for most systems of biological interest. Moreover, the models usually consist of a large number of variables, are strongly nonlinear, and include different time-scales, which makes them difficult to handle both mathematically and computationally. This has motivated the interest in qualitative models which, from incomplete knowledge of the system, are able to provide a coarse-grained picture of its dynamics.

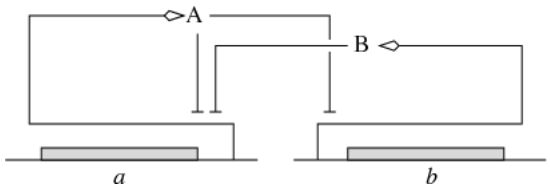
A variety of qualitative modeling formalisms have been introduced over the past decades. Boolean or logical models, which describe gene regulatory and signalling networks as discrete-time finite-state transition systems, are probably most widely used. The dynamics of these systems are governed by logical functions representing the regulatory interactions between the genes and other components of the system. IBIS has focused on a related, hybrid formalism that embeds the logical functions describing regulatory interactions into an ODE formalism, giving rise to so-called piecewise-linear differential equations (PLDEs, Figure 2). The use of logical functions allows the qualitative dynamics of the PLDE models to be analyzed, even in high-dimensional systems. In particular, the qualitative dynamics can be represented by means of a so-called state transition graph, where the states correspond to (hyperrectangular) regions in the state space and transitions between states arise from solutions entering one region from another.

First proposed by Leon Glass and Stuart Kauffman in the early seventies, the mathematical analysis of PLDE models has been the subject of active research for more than four decades. IBIS has made contributions on the mathematical level, in collaboration with the BIOCORE and BIPOP project-teams, notably for solving problems induced by discontinuities in the dynamics of the system at the boundaries between regions, where the logical functions may abruptly switch from one discrete value to another, corresponding to the (in)activation of a gene. In addition, many efforts have gone into the development of the computer tool GENETIC NETWORK ANALYZER (GNA) and its applications to the analysis of the qualitative dynamics of a variety of regulatory networks in microorganisms. Some of the methodological work underlying GNA, notably the development of analysis tools based on temporal logics and model checking, which was carried out with the Inria project-teams CONVEX (ex-VASY) and POP-ART, has implications beyond PLDE models as they apply to logical and other qualitative models as well.

3.2. Inference of gene regulatory networks from time-series data

Participants: Eugenio Cinquemani [Correspondent], Johannes Geiselmann, Hidde de Jong, Stéphan Lacour, Aline Marguet, Michel Page, Corinne Pinel, Delphine Ropers.

Measurements of the transcriptome of a bacterial cell by means of DNA microarrays, RNA sequencing, and other technologies have yielded huge amounts of data on the state of the transcriptional program in different growth conditions and genetic backgrounds, across different time-points in an experiment. The information on the time-varying state of the cell thus obtained has fueled the development of methods for inferring regulatory interactions between genes. In essence, these methods try to explain the observed variation in the activity of one gene in terms of the variation in activity of other genes. A large number of inference methods have been proposed in the literature and have been successful in a variety of applications, although a number of difficult problems remain.



$$\dot{x}_a = \kappa_a \, s^-(x_a, \theta_a^2) \, s^-(x_b, \theta_b) - \gamma_a \, x_a$$
$$\dot{x}_b = \kappa_b \, s^-(x_a, \theta_a^1) - \gamma_b \, x_b$$
$$s^+(x, \theta) = \begin{cases} 1, & \text{if } x > \theta \\ 0, & \text{if } x < \theta \end{cases}$$
$$s^-(x, \theta) = 1 - s^+(x, \theta)$$

Figure 2. (Left) Example of a gene regulatory network of two genes (a and b), each of which codes for a regulatory protein (A and B). Protein B inhibits the expression of gene a, while protein A inhibits the expression of gene b and its own gene. (Right) PLDE model corresponding to the network in (a). Protein A is synthesized at a rate κ_a, if and only if the concentration of protein A is below its threshold θ_a^2 ($x_a < \theta_a^2$) and the concentration of protein B below its threshold θ_b ($x_b < \theta_b$). The degradation of protein A occurs at a rate proportional to the concentration of the protein itself ($\gamma_a \, x_a$).

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) and other fluorescent and luminescent reporter proteins, provide an excellent means to measure the activity of a gene *in vivo* and in real time (Figure 3). The underlying principle of the technology is to fuse the promoter region and possibly (part of) the coding region of a gene of interest to a reporter gene. The expression of the reporter gene generates a visible signal (fluorescence or luminescence) that is easy to capture and reflects the expression of a gene of interest. The interest of the reporter systems is further enhanced when they are applied in mutant strains or combined with expression vectors that allow the controlled induction of any particular gene, or the degradation of its product, at a precise moment during the time-course of the experiment. This makes it possible to perturb the network dynamics in a variety of ways, thus obtaining precious information for network inference.

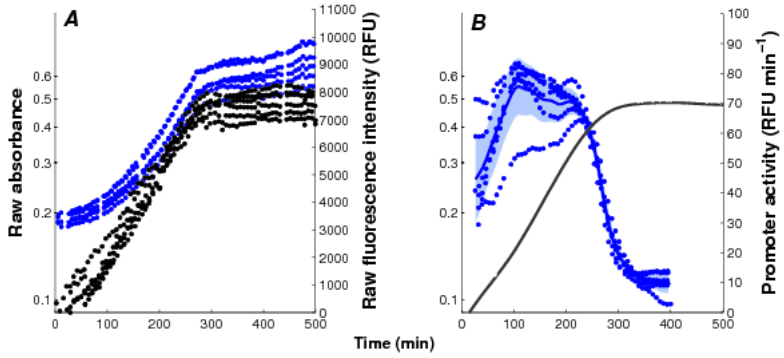


Figure 3. Monitoring of bacterial gene expression in vivo using fluorescent reporter genes (Stefan et al., PLoS Computational Biology, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with E. coli cells, focusing on the expression of the motility gene tar in a mutant background. A: Absorbance (•, black) and fluorescence (•, blue) data, corrected for background intensities, obtained with the ΔcpxR strain transformed with the ptar-gfp reporter plasmid and grown in M9 with glucose. B: Activity of the tar promoter, computed from the primary data. The solid black line corresponds to the mean of 6 replicate absorbance measurements and the shaded blue region to the mean of the promoter activities ± twice the standard error of the mean.

The specific niche of IBIS in the field of network inference has been the development and application of genome engineering techniques for constructing the reporter and perturbation systems described above, as well as the use of reporter gene data for the reconstruction of gene regulation functions. We have developed an experimental pipeline that resolves most technical difficulties in the generation of reproducible time-series measurements on the population level. The pipeline comes with data analysis software that converts the primary data into measurements of time-varying promoter activities. In addition, for measuring gene expression on the single-cell level by means of microfluidics and time-lapse fluorescence microscopy, we have established collaborations with groups in Grenoble and Paris. The data thus obtained can be exploited for the structural and parametric identification of gene regulatory networks, for which methods with a solid mathematical foundation are developed, in collaboration with colleagues at ETH Zürich and EPF Lausanne (Switzerland). The vertical integration of the network inference process, from the construction of the biological material to the data analysis and inference methods, has the advantage that it allows the experimental design to be precisely tuned to the identification requirements.

3.3. Analysis of integrated metabolic and gene regulatory networks

Participants: Eugenio Cinquemani, Hidde de Jong, Thibault Etienne, Johannes Geiselmann, Stéphan Lacour, Yves Markowicz, Marco Mauri, Michel Page, Corinne Pinel, Delphine Ropers [Correspondent].

The response of bacteria to changes in their environment involves responses on several different levels, from the redistribution of metabolic fluxes and the adjustment of metabolic pools to changes in gene expression. In order to fully understand the mechanisms driving the adaptive response of bacteria, as mentioned above, we need to analyze the interactions between metabolism and gene expression. While often studied in isolation, gene regulatory networks and metabolic networks are closely intertwined. Genes code for enzymes which control metabolic fluxes, while the accumulation or depletion of metabolites may affect the activity of transcription factors and thus the expression of enzyme-encoding genes.

The fundamental principles underlying the interactions between gene expressions and metabolism are far from being understood today. From a biological point of view, the problem is quite challenging, as metabolism and gene expression are dynamic processes evolving on different time-scales and governed by different types of kinetics. Moreover, gene expression and metabolism are measured by different experimental methods generating heterogeneous, and often noisy and incomplete data sets. From a modeling point of view, difficult methodological problems concerned with the reduction and calibration of complex nonlinear models need to be addressed.

Most of the work carried out within the IBIS project-team specifically addressed the analysis of integrated metabolic and gene regulatory networks in the context of *E. coli* carbon metabolism (Figure 4). While an enormous amount of data has accumulated on this model system, the complexity of the regulatory mechanisms and the difficulty to precisely control experimental conditions during growth transitions leave many essential questions open, such as the physiological role and the relative importance of mechanisms on different levels of regulation (transcription factors, metabolic effectors, global physiological parameters, ...). We are interested in the elaboration of novel biological concepts and accompanying mathematical methods to grasp the nature of the interactions between metabolism and gene expression, and thus better understand the overall functioning of the system. Moreover, we have worked on the development of methods for solving what is probably the hardest problem when quantifying the interactions between metabolism and gene expression: the estimation of parameters from hetereogeneous and noisy high-throughput data. These problems are tackled in collaboration with experimental groups at Inra/INSA Toulouse and CEA Grenoble, which have complementary experimental competences (proteomics, metabolomics) and biological expertise.

3.4. Natural and engineered control of growth and gene expression

Participants: Célia Boyat, Eugenio Cinquemani, Johannes Geiselmann [Correspondent], Hidde de Jong [Correspondent], Stéphan Lacour, Marco Mauri, Tamas Muszbek, Michel Page, Antrea Pavlou, Delphine Ropers, Maaike Sangster.

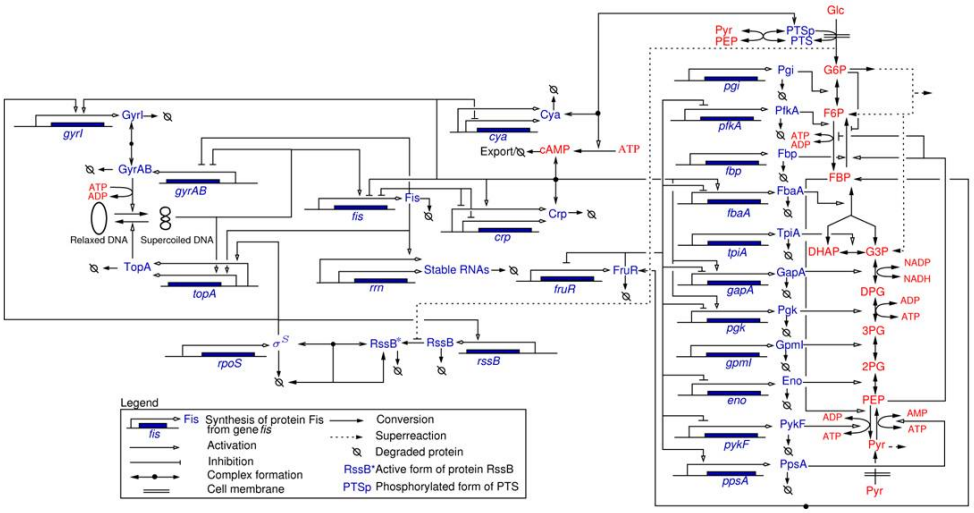


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the PTS system, via the phosphorylated and non-phosphorylated form of its enzymes (represented by PTSp and PTS, respectively). The pentose-phosphate pathway (PPP) is not explicitly described but we take into account that a small pool of G6P escapes the upper part of glycolysis. At the level of the global regulators the network includes the control of the DNA supercoiling level, the accumulation of the sigma factor RpoS and the Crp·cAMP complex, and the regulatory role exerted by the fructose repressor FruR.

The adaptation of bacterial physiology to changes in the environment, involving changes in the growth rate and a reorganization of gene expression, is fundamentally a resource allocation problem. It notably poses the question how microorganisms redistribute their protein synthesis capacity over different cellular functions when confronted with an environmental challenge. Assuming that resource allocation in microorganisms has been optimized through evolution, for example to allow maximal growth in a variety of environments, this question can be fruitfully formulated as an optimal control problem. We have developed such an optimal control perspective, focusing on the dynamical adaptation of growth and gene expression in response to envrionmental changes, in close collaboration with the BIOCORE project-team.

A complementary perspective consists in the use of control-theoretical approaches to modify the functioning of a bacterial cell towards a user-defined objective, by rewiring and selectively perturbing its regulatory networks. The question how regulatory networks in microorganisms can be externally controlled using engineering approaches has a long history in biotechnology and is receiving much attention in the emerging field of synthetic biology. Within a number of on-going projects, IBIS is focusing on two different questions. The first concerns the development of open-loop and closed-loop growth-rate controllers of bacterial cells for both fundamental research and biotechnological applications (Figure 5). Second, we are working on the development of methods for the real-time control of the expression of heterologous proteins in communities of interacting bacterial populations. The above projects involve collaborations with, among others, the Inria project-teams LIFEWARE (INBIO), BIOCORE, and McTAO as well as with a biophysics group at Univ Paris Descartes and a mathematical modeling group at INRA Jouy-en-Josas.

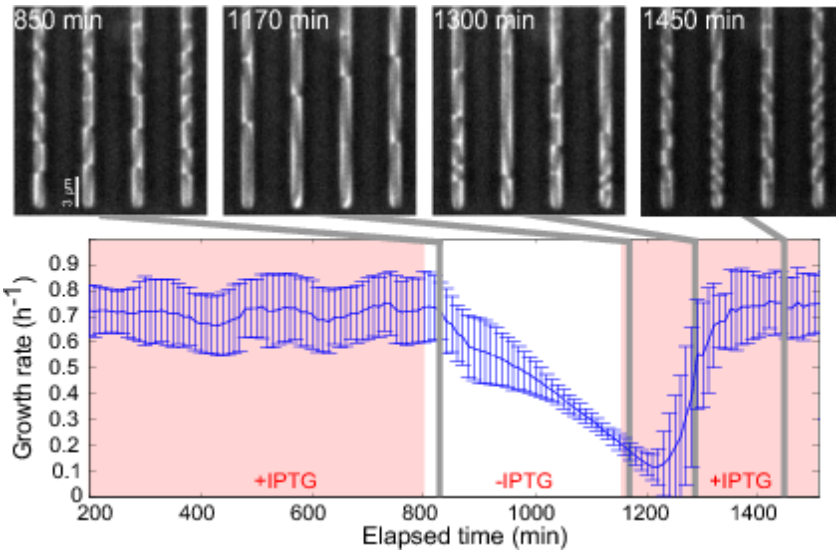


Figure 5. Growth arrest by external control of the gene expression machinery (Izard, Gomez Balderas et al., Molecular Systems Biology, 11:840, 2015). An E. coli strain in which an essential component of the gene expression machinery, the $\beta\beta'$ subunits of RNA polymerase, was put under the control of an externally-supplied inducer (IPTG), was grown in a microfluidics device and phase-contrast images were acquired every 10 min. The cells were grown in minimal medium with glucose, initially in the presence of 1 mM IPTG. 6 h after removing IPTG from the medium, the growth rate slows down and cells are elongated. About 100 min after adding back 1 mM IPTG into the medium, the elongated cells divide and resume normal growth. The growth rates in the plot are the (weighted) mean of the growth rates of 100 individual cells. The error bars correspond to $\pm$ one standard deviation. The results of the experiment show that the growth rate of a bacterial can be switched off in a reversible manner by an external inducer, based on the reengineering of the natural control of the expression of RNA polymerase.

LEMON Project-Team

3. Research Program

3.1. Foreword

The team has three main scientific objectives. The first is to develop new models and advanced mathematical methods for inland flow processes. The second is to investigate the derivation and use of coupled models for marine and coastal processes (mainly hydrodynamics, but not only). The third is to develop theoretical methods to be used in the mathematical models serving the first two objectives. As mentioned above, the targeted applications cover PDE models and related extreme events using a hierarchy of models of increasing complexity. LEMON members also contribute to research projects that are not in the core of the team topics and that correspond to external collaborations: they are mentioned in the fourth section below.
In every section, people involved in the project are listed in alphabetical order, except for the first one (underlined) which corresponds to the leading scientist on the corresponding objective.

3.2. Inland flow processes

3.2.1. Shallow water models with porosity

3.2.1.1. State of the Art

Simulating urban floods and free surface flows in wetlands requires considerable computational power. Two-dimensional shallow water models are needed. Capturing the relevant hydraulic detail often requires computational cell sizes smaller than one meter. For instance, meshing a complete urban area with a sufficient accuracy would require 10^6 to 10^8 cells, and simulating one second often requires several CPU seconds. This makes the use of such model for crisis management impossible. Similar issues arise when modelling wetlands and coastal lagoons, where large areas are often connected by an overwhelming number of narrow channels, obstructed by vegetation and a strongly variable bathymetry. Describing such channels with the level of detail required in a 2D model is impracticable. A new generation of models overcoming this issue has emerged over the last 20 years: porosity-based shallow water models. They are obtained by averaging the two-dimensional shallow water equations over large areas containing both water and a solid phase [29]. The size of a computational cell can be increased by a factor 10 to 50 compared to a 2D shallow water model, with CPU times reduced by 2 to 3 orders of magnitude [48]. While the research on porosity-based shallow water models has accelerated over the past decade [43], [59], [62], [39], [38], [48], [73], [74], [68], [69], a number of research issues remain pending.

3.2.1.2. Four year research objectives

The research objectives are (i) to improve the upscaling of the flux and source term models to be embedded in porosity shallow water models, (ii) to validate these models against laboratory and in situ measurements. Improving the upscaled flux and source term models for urban applications requires that description of anisotropy in porosity models be improved to account for the preferential flows induced by building and street alignment. The description of the porosity embedded in the most widespread porosity approach, the so-called Integral Porosity model [59], [41], has been shown to provide an incomplete description of the connectivity properties of the urban medium. Firstly, the governing equations are strongly mesh-dependent because of consistency issues [41]. Secondly, the flux and source term models fail to reproduce the alignment with the main street axes in a number of situations [40]. Another path for improvement concerns the upscaling of obstacle-induced drag terms in the presence of complex geometries. Recent upscaling research results obtained by the LEMON team in collaboration with Tour du Valat suggest that the effects of microtopography on the flow cannot be upscaled using "classical" equation-of-state approaches, as done in most hydraulic models. A totally different approach must be proposed. The next four years will be devoted to the development and validation of improved flux and source term closures in the presence of strongly anisotropic urban geometries

and in the presence of strongly variable topography. Validation will involve not only the comparison of porosity model outputs with refined flow simulation results, but also the validation against experimental data sets. No experimental data set allowing for a sound validation of flux closures in porosity models can be found in the literature. Laboratory experiments will be developed specifically in view of the validation of porosity models. Such experiments will be set up and carried out in collaboration with the Université Catholique de Louvain (UCL), that has an excellent track record in experimental hydraulics and the development of flow monitoring and data acquisition equipment. These activities will take place in the framework of the PoroCity Associate International Laboratory (see next paragraph).

3.2.1.3. People

Vincent Guinot, Carole Delenne, Pascal Finaud-Guyot, Antoine Rousseau.

3.2.1.4. External collaborations

- Tour du Valat (O. Boutron): the partnership with TdV focuses on the development and application of depth-dependent porosity models to the simulation of coastal lagoons, where the bathymetry and geometry is too complex to be represented using refined flow models.
- University of California Irvine (B. Sanders): the collaboration with UCI started in 2014 with research on the representation of urban anisotropic features in integral porosity models [48]. It has led to the development of the Dual Integral Porosity model [42]. Ongoing research focuses on improved representations of urban anisotropy in urban floods modelling.
- Université Catholique de Louvain - UCL (S. Soares-Frazão): UCL is one of the few places with experimental facilities allowing for the systematic, detailed validation of porosity models. The collaboration with UCL started in 2005 and will continue with the PoroCity Associate International Laboratory proposal. In this proposal, a four year research program is set up for the validation, development and parametrization of shallow water models with porosity.
- Luxembourg Institute of Technology (R. Hostache): the collaboration with LIST started in 2018 with the project CASCADE funded by the Fond National de la Recherche du Luxembourg, and the co-direction of Vita Ayoub. The depth-dependant porosity model is applied to simulate the flooding of the Severn river (UK).

3.2.2. Forcing

3.2.2.1. State of the Art

Reproducing optimally realistic spatio-temporal rainfall fields is of salient importance to the forcing of hydrodynamic models. This challenging task requires combining intense, usual and dry weather events. Far from being straightforward, this combination of extreme and non-extreme scenarii requires a realistic modelling of the transitions between normal and extreme periods. [52] have proposed in a univariate framework a statistical model that can serve as a generator and that takes into account low, moderate and intense precipitation. In the same vein, [70] developed a bivariate model. However, its extension to a spatial framework remains a challenge. Existing spatial precipitation stochastic generators are generally based on Gaussian spatial processes [15], [50], that are not adapted to generate extreme rainfall events. Recent advances in spatio-temporal extremes modelling based on generalized Pareto processes [32], [65] and semi-parametric simulation techniques [21] are very promising and could form the base for relevant developments in our framework.

3.2.2.2. Four year research objectives

The purpose is to develop stochastic methods for the simulation of realistic spatio-temporal processes integrating extreme events. Two steps are identified. The first one is about the simulation of extreme events and the second one concerns the combination of extreme and non extreme events in order to build complete, realistic precipitations time series. As far as the first step is concerned, a first task is to understand and to model the space-time structure of hydrological extremes such as those observed in the French Mediterranean basin, that is known for its intense rainfall events (Cevenol episodes), which have recently received increased attention. We will propose modelling approaches based on the exceedance, which allows the simulated fields to be interpreted as events. Parametric, semi-parametric and non-parametric approaches are currently under

consideration. They would allow a number of scientific locks to be removed. Examples of such locks are e.g. accounting for the temporal dimension and for various dependence structures (asymptotic dependence or asymptotic independence possibly depending on the dimension and/or the distance considered). Methodological aspects are detailed in Section 3.4.1 . The second step, which is not straightforward, consists in combining different spatio-temporal simulations in order to help to ultimately develop a stochastic precipitation generator capable of producing full precipitation fields, including dry and non-extreme wet periods.

3.2.2.3. People

Gwladys Toulemonde, Carole Delenne, Vincent Guinot.

3.2.2.4. External collaborations

The Cerise (2016-2018) and Fraise (2019-2021) projects (see 8.2), led by Gwladys Toulemonde, are funded by the action MANU (MAthematical and Numerical methods) of the CNRS LEFE program [0]. Among others, they aim to propose methods for simulating scenarii integrating spatio-temporal extremes fields with a possible asymptotic independence for impact studies in environmental sciences. Among the members of this project, Jean-Noel Bacro (IMAG, UM), Carlo Gaetan (DAIS, Italy) and Thomas Opitz (BioSP, MIA, INRA) are involved in the first step as identified in the research objectives of the present sub-section. Denis Allard (BioSP, MIA, INRA), Julie Carreau (IRD, HSM) and Philippe Naveau (CNRS, LSCE) will be involved in the second one.

3.2.3. Parametrization of shallow water models with porosity

3.2.3.1. State of the Art

Numerical modelling requires data acquisition, both for model validation and for parameter assessment. Model benchmarking against laboratory experiments is an essential step and is an integral part of the team's strategy. However, scale model experiments may have several drawbacks: (i) experiments are very expensive and extremely time-consuming, (ii) experiments cannot always be replicated, and measurement have precision and reliability limitations, (iii) dimensional similarity (in terms of geometry and flow characteristic variables such as Froude or Reynolds numbers) cannot always be preserved.

An ideal way to obtain data would be to carry out in situ measurements. But this would be too costly at the scale of studied systems, not to mention the fact that field may become impracticable during flood periods.

Geographical and remote sensing data are becoming widely available with high spatial and temporal resolutions. Several recent studies have shown that flood extends can be extracted from optical or radar images [35], for example: to characterize the flood dynamics of great rivers [53], to monitor temporary ponds [63], but also to calibrate hydrodynamics models and assess roughness parameters (e.g. [72].

Upscaled models developed in LEMON (see 3.2.1) embed new parameters that reflect the statistical properties of the medium geometry and the subgrid topography. New methods are thus to be developed to characterize such properties from remote sensing and geographical data.

3.2.3.2. Four year research objectives

This research line consists in deriving methods and algorithms for the determination of upscaled model parameters from geodata.

For applications in urban areas, it is intended to extract information on the porosity parameters from National geographical survey databases largely available in developed countries. Such databases usually incorporate separate layers for roads, buildings, parking lots, yards, etc. Most of the information is stored in vector form, which can be expected to make the treatment of urban anisotropic properties easier than with a raster format. In developing countries, data is made increasingly available over the world thanks to crowdsourcing (e.g. OpenStreetMap) the required level of detail sometimes not available in vector format, especially in suburban areas, where lawns, parks and other vegetated areas, that may also contribute to flood propagation and storage, are not always mapped. In this context, the necessary information can be extracted from aerial and/or satellite images, that are widely available and the spatial resolution of which improves constantly, using supervised classification approaches.

[0] Les Enveloppes Fluides et l'Environnement

For applications in great rivers the main objective is to develop an efficient framework for optimally integrating remote sensing derived flood information to compensate the lack of observation related to riverbed bathymetry and river discharge. The effective integration of such remote sensing-derived flood information into hydraulic models remains a critical issue. In partnership with R. Hostache (LIST), we will investigate new ways for making use of SEO data (i.e. flooded areas and water level estimates derived from SAR data collections) for retrieving uncertain model parameters and boundary conditions. The method will be developed and validated using synthetically generated data sets as well as real-event data retrieved from the European Space Agency's archives. Extensive testing will be carried out in a number of high magnitude events recorded over the Severn (United Kingdom) and Zambezi (Mozambique) floodplain areas.

In wetlands applications, connectivity between different ponds is highly dependent on the free surface elevation, thus conditioning the presence of a flow. Characterizing such connectivity requires that topographical variations be known with high accuracy. Despite the increased availability of direct topographic measurements from LiDARS on riverine systems, data collection remains costly when wide areas are involved. Data acquisition may also be difficult when poorly accessible areas are dealt with. If the amount of topographic points is limited, information on elevation contour lines can be easily extracted from the flood dynamics visible in simple SAR or optical images. A challenge is thus to use such data in order to estimate continuous topography on the floodplain combining topographic sampling points and located contour lines the levels of which are unknown or uncertain.

3.2.3.3. People

Carole Delenne, Vincent Guinot, Antoine Rousseau, Pascal Finaud-Guyot

3.2.3.4. External collaborations

- A first attempt for topography reconstruction in wetlands was done in collaboration with J.-S. Bailly (LISAH) in 2016 [30]. It is intended to reactivate this topic in the coming years.
- Porosity model calibration for application on great rivers will be done in the framework of CASCADE project in collaboration with R. Hostache (LIST).
- A collaboration started with the LISAH laboratory to investigate the feasibility of depth-dependent porosity laws reconstruction over cultivates areas. LISAH personel involved: D. Feurer, D. Raclot.

3.3. Marine and coastal systems

3.3.1. Multi-scale ocean modelling

The expertise of LEMON in this scientific domain is more in the introduction and analysis of new boundary conditions for ocean modelling systems, that can be tested on academical home-designed test cases. This is in the core of Antoine Rousseau's contributions over the past years. The real implementation, within operational ocean models, has to be done thanks to external collaborations which have already started with LEMON (see below).

3.3.1.1. State of the Art

In physical oceanography, all operational models - regardless of the scale they apply to - are derived from the complete equations of geophysical fluid dynamics. Depending on the considered process properties (nonlinearity, scale) and the available computational power, the original equations are adapted with some simplifying hypotheses. The reader can refer to [58], [51] for a hierarchical presentation of such models.
In the nearshore area, the hydrostatic approximation that is used is most large scales models (high sea) cannot be used without a massive loss of accuracy. In particular, shallow water models are inappropriate to describe the physical processes that occur in this zone (see Figure 1). This is why Boussinesq-type models are prefered: see [49]. They embed dispersive terms that allow for shoaling and other bathymetry effects. Since the pioneering works of Green and Naghdi (see [36]), numerous theoretical and numerical studies have been delivered by the "mathematical oceanography" community, more specifically in France (see the works of Lannes, Marche, Sainte-Marie, Bresch, etc.). The corresponding numerical models (BOSZ, WaveBox) must thus be integrated in any reasonable nearshore modelling platform.

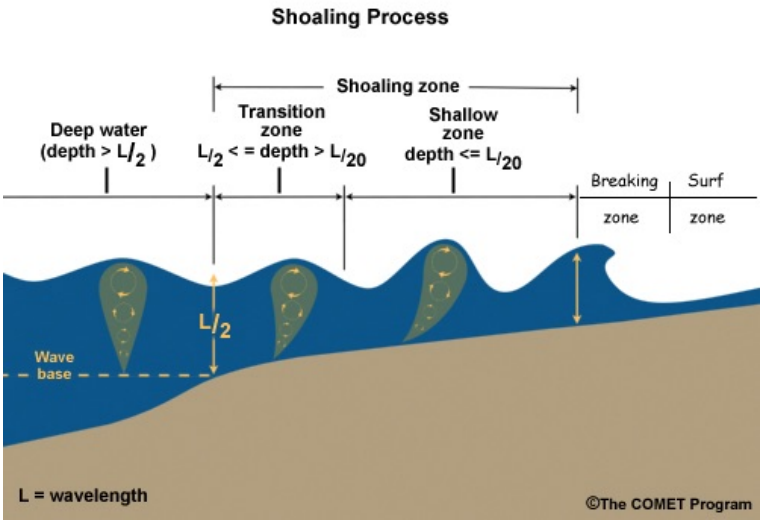


Figure 1. Deep sea, shoaling, and breaking zones.

However, these models cannot simply replace all previous models everywhere in the ocean: dispersive models are useless away from the shore and it is known that wave breaking cannot be simulated using Boussinesq-type equations. Hence the need to couple these models with others. Some work has been done in this direction with a multi-level nesting using software packages such as ROMS, but to the best of our knowledge, all the "boxes" rely on the same governing equations with different grid resolutions. A real coupling between different models is a more difficult task since different models may have different mathematical properties, as shown in the work by Eric Blayo and Antoine Rousseau on shallow water modelling (see [17]).

3.3.1.2. Four year research objectives

Starting from the knowledge acquired in the collaboration with Eric Blayo on model coupling using domain decomposition techniques, our ambition is to propose theoretical and numerical tools in order to incorporate nearshore ocean models into large complex systems including several space and time scales. Two complementary research directions are considered:

- **Dispersive *vs* non-dispersive shallow water models**. As depicted in Figure 1 above, Boussinesq-type models (embedding dispersive effects) should be used in the so-called shoaling zone. The coupling with classical deep-sea / shallow water models has to be done such that all the processes in Figure 1 are correctly modelled (by different equations), with a reduced numerical cost. As a first guess, we think that Schwarz-type methods (widely used by the DDM community) could be good candidates, in particular when the interface locations are well-known. Moving interfaces (depending on the flow, the bathymetry and naturally the wind and all external forcings) is a more challenging objective that will be tackled after the first step (known interface) is achieved.
- **spectral *vs* time-domain models**. In the context of mathematical modelling and numerical simulation for the marine energy, we want to build a coupled numerical model that would be able to simulate wave propagation in domains covering both off-shore regions, where spectral models are used, and nearshore regions, better described by nonlinear dispersive (Boussinesq-type) models.
  While spectral models work with a statistical and phase-averaged description of the waves, solving the evolution of its energy spectrum, Boussinesq-type models are phase-resolving and solves nonlinear dispersive shallow water equations for physical variables (surface elevation and velocity)

in the time domain. Furthermore, the time and space scales are very different: they are much larger in the case of spectral models, which justifies their use for modelling off-shore propagation over large time frames. Moreover, important small scale phenomena in nearshore areas are better captured by Boussinesq models, in which the time step is limited by the CFL condition.
From a mathematical and modelling point of view, this task mainly consists in working on the boundary conditions of each model, managing the simultaneous use of spectral and time series data, while studying transparent boundary conditions for the models and developing domain decomposition approaches to improve the exchange of information.

3.3.1.3. People

Antoine Rousseau, Joao Guilherme Caldas Steinstraesser

3.3.1.4. External collaborations

- **Eric Blayo** is the former scientific leader of team MOISE in Grenoble, where Antoine Rousseau was first recruited. Eric Blayo and Antoine Rousseau have co-advised 3 PhDs and continue to work together on coupling methods in hydrodynamics, especially in the framework of the COMODO ANR network.
- **Fabien Marche** (at IMAG, Montpellier, currently on leave in Bordeaux) is an expert in numerical modelling and analysis of Boussinesq-type models. He is the principal investigator of the WaveBox software project, to be embedded in the national scale Uhaina initiative.
- In the framework of its collaboration with **MERIC**, Antoine Rousseau and Joao Guilherme Caldas Steinstraesser collaborate with the consortium DiMe (ANR-FEM project), and more particularly with Jean-François Filipot ans Volker Roeber for the coupling of spectral and time-domain methods.

3.3.2. Data-model interactions

3.3.2.1. State of the Art

An alternative to direct observations is the chaining of numerical models, which for instance represent the physic from offshore to coastal areas. Typically, output data from atmospheric and ocean circulation models are used as forcings for a wave model, which in turn feeds a littoral model. In the case of extreme events, their numerical simulation from physical models is generally unreachable. This is due to a lack of knowledge on boundary conditions and on their physical reliability for such extreme quantities. Based on numerical simulated data, an alternative is to use statistical approaches. [21] proposed such an approach. They first produced and studied a 52-year hindcast using the WW3 wave model [19], [22], [20], [66]. Then stemming from parts of the original work of [18], [37], [32], [21] proposed a semi-parametric approach which aims to simulate extreme space-time waves processes to, in turn, force a littoral hazard model. Nevertheless their approach allows only a very small number of scenarii to be simulated.

3.3.2.2. Four year research objectives

A first objective is to establish the link between the simulation approach proposed by [21] and the Pareto Processes [32]. This will allow the work of [21] to be generalized, thus opening up the possibility of generating an infinity of extreme scenarii. While continuing to favor the semi- or non-parametric approaches made possible by the access to high spatial resolution calculations, we will try to capture the strength of potentially decreasing extremal dependence when moving towards higher values, which requires the development of models that allow for so-called asymptotic independence.

3.3.2.3. People

Gwladys Toulemonde, Fátima Palacios Rodríguez, Antoine Rousseau

3.3.2.4. External collaborations

- since late 2019, LEMON has started a collaboration with IRT Saint-Exupéry on the hybridization of models and large amounts of data for the modelling of urban floods
- The collaboration with Romain Chailan (IMAG, UM, CNRS) and Frédéric Bouchette (Geosciences, UM) started in 2012 during the PhD of Romain entitled Application of scientific computing and statistical analysis to address coastal hazards.
- During her post doctoral position, Fátima Palacios Rodríguez with her co-advisors will considered a generalization of the proposed simulation method by [21].

3.4. Methodological developments

In addition to the application-driven sections, the team also works on the following theoretical questions. They are clearly connected to the abovementioned scientific issues but do not correspond to a specific application or process.

3.4.1. Stochastic models for extreme events

3.4.1.1. State of the Art

Max-stable random fields [61], [60], [46], [26], [54] are the natural limit models for spatial maximum data and have spawned a very rich literature. An overview of typical approaches to modelling maxima is due to [28]. Physical interpretation of simulated data from such models can be discussed. An alternative to the max-stable framework are models for threshold exceedances. Processes called GPD processes, which appear as a generalization of the univariate formalism of the high thresholds exceeding a threshold based on the GPD, have been proposed [32], [65]. Strong advantages of these thresholding techniques are their capability to exploit more information from the data and explicitly model the original event data. However, the asymptotic dependence stability in these limiting processes for maximum and threshold exceedance tends to be overly restrictive when asymptotic dependence strength decreases at high levels and may ultimately vanish in the case of asymptotic independence. Such behaviours appear to be characteristic for many real-world data sets such as precipitation fields [27], [64]. This has motivated the development of more flexible dependence models such as max-mixtures of max-stable and asymptotically independent processes [71], [13] for maxima data, and Gaussian scale mixture processes [55], [45] for threshold exceedances. These models can accommodate asymptotic dependence, asymptotic independence and Gaussian dependence with a smooth transition. Extreme events also generally present a temporal dependence [67] . Developing flexible space-time models for extremes is crucial for characterizing the temporal persistence of extreme events spanning several time steps; such models are important for short-term prediction in applications such as the forecasting of wind power and for extreme event scenario generators providing inputs to impact models, for instance in hydrology and agriculture. Currently, only few models are available from the statistical literature (see for instance [24], [25], [44]) and remain difficult to interpret.

3.4.1.2. Four year research objectives

The objective is to extend state-of-the-art methodology with respect to three important aspects: 1) adapting well-studied spatial modelling techniques for extreme events based on asymptotically justified models for threshold exceedances to the space-time setup; 2) replacing restrictive parametric dependence modelling by semiparametric or nonparametric approaches; 3) proposing more flexible spatial models in terms of asymmetry or in terms of dependence. This means being able to capture the strength of potentially decreasing extremal dependence when moving towards higher values, which requires developing models that allow for so-called asymptotic independence.

3.4.1.3. People

Gwladys Toulemonde, Fátima Palacios Rodríguez

3.4.1.4. External collaborations

In a natural way, the Cerise and Fraise project members are the main collaborators for developing and studying new stochastic models for extremes.

- More specifically, research with Jean-Noel Bacro (IMAG, UM), Carlo Gaetan (DAIS, Italy) and Thomas Opitz (BioSP, MIA, INRA) focuses on relaxing dependence hypothesis.
- The asymmetry issue and generalization of some Copula-based models are studied with Julie Carreau (IRD, HydroSciences, UM).

3.4.2. Integrating heterogeneous data

3.4.2.1. State of the Art

Assuming that a given hydrodynamic models is deemed to perform satisfactorily, this is far from being sufficient for its practical application. Accurate information is required concerning the overall geometry of the area under study and model parametrization is a necessary step towards the operational use. When large areas are considered, data acquisition may turn out prohibitive in terms of cost and time, not to mention the fact that information is sometimes not accessible directly on the field. To give but one example, how can the roughness of an underground sewer pipe be measured? A strategy should be established to benefit from all the possible sources of information in order to gather data into a geographical database, along with confidence indexes.

The assumption is made that even hardly accessible information often exists. This stems from the increasing availability of remote-sensing data, to the crowd-sourcing of geographical databases, including the inexhaustible source of information provided by the Internet. However, information remains quite fragmented and stored in various formats: images, vector shapes, texts, etc.

This path of research begun with the Cart'Eaux project (2015-2018), that aims to produce regular and complete mapping of urban wastewater system. Contrary to drinkable water networks, the knowledge of sewer pipe location is not straightforward, even in developed countries. Over the past century, it was common practice for public service providers to install, operate and repair their networks separately [57]. Now local authorities are confronted with the task of combining data produced by different parts, having distinct formats, variable precision and granularity [23].

3.4.2.2. Four year research objectives

The overall objective of this research line is to develop methodologies to gather various types of data in the aim of producing an accurate mapping of the studied systems for hydrodynamics models.

Concerning wastewater networks, the methodology applied consists in inferring the shape of the network from a partial dataset of manhole covers that can be detected from aerial images [56]. Since manhole covers positions are expected to be known with low accuracy (positional uncertainty, detection errors), a stochastic algorithm is set up to provide a set of probable network geometries [4]. As more information is required for hydraulic modelling than the simple mapping of the network (slopes, diameters, materials, etc.), text mining techniques such as used in [47] are particularly interesting to extract characteristics from data posted on the Web or available through governmental or specific databases. Using an appropriate keyword list, thematic entities are identified and linked to the surrounding spatial and temporal entities in order to ease the burden of data collection. It is clear at this stage that obtaining numerical values on specific pipes will be challenging. Thus, when no information is found, decision rules will be used to assign acceptable numerical values to enable the final hydraulic modelling.

In any case, the confidence associated to each piece of data, be it directly measured or reached from a roundabout route, should be assessed and taken into account in the modelling process. This can be done by generating a set of probable inputs (geometry, boundary conditions, forcing, etc.) yielding simulation results along with the associated uncertainty.

Combining heterogeneous data for a better knowledge of studied systems raises the question of data fusion. What is the reality when contradictory information is collected from different sources? Dealing with spatial information, offset are quite frequent between different geographical data layers; pattern comparison approaches should be developed to judge whether two pieces of information represented by two elements close to each other are in reality identical, complementary, or contradictory.

3.4.2.3. People

Carole Delenne, Vincent Guinot, Antoine Rousseau, Gwladys Toulemonde

3.4.2.4. External collaborations

The Cart'Eaux project has been a lever to develop a collaboration with Berger-Levrault company and several multidisciplinary collaborations for image treatment (LIRMM), text analysis (LIRMM and TETIS) and network cartography (LISAH, IFSTTAR).

- The MeDo project lead by N. Chahinian (HSM) in collaboration with linguists of UMR Praxiling, uses data mining and text analysis approaches to retrieve information on wastewater networks from the Web. Carole Delenne has a slight implication in this project, as domain expert to guide the text annotations and for the uncertainties definition and representation in the mapping of the data collected.
- Concerning geographical data fusion for the wastewater network cartograhpy, the Phd thesis of Yassine Bel-Ghaddar has been funded by the French Association of Research and Technology (ANRT) in collaboration with Berger-Levrault company and in co-direction with A. Begdouri (LSIA Fès, Morocco).

3.4.3. Numerical methods for porosity models

3.4.3.1. State of the Art

Porosity-based shallow water models are governed by hyperbolic systems of conservation laws. The most widespread method used to solve such systems is the finite volume approach. The fluxes are computed by solving Riemann problems at the cell interfaces. This requires that the wave propagation properties stemming from the governing equations be known with sufficient accuracy. Most porosity models, however, are governed by non-standard hyperbolic systems.

Firstly, the most recently developed DIP models include a momentum source term involving the divergence of the momentum fluxes [42]. This source term is not active in all situations but takes effect only when positive waves are involved [39], [40]. The consequence is a discontinuous flux tensor and discontinuous wave propagation properties. The consequences of this on the existence and uniqueness of solutions to initial value problems (especially the Riemann problem) are not known, or are the consequences on the accuracy of the numerical methods used to solve this new type of equations.

Secondly, most applications of these models involve anisotropic porosity fields [48], [59]. Such anisotropy can be modelled using 2×2 porosity tensors, with principal directions that are not aligned with those of the Riemann problems in two dimensions of space. The solution of such Riemann problems has not been investigated yet. Moreover, the governing equations not being invariant by rotation, their solution on unstructured grids is not straightforward.

Thirdly, the Riemann-based, finite volume solution of the governing equations require that the Riemann problem be solved in the presence of a porosity discontinuity. While recent work [31] has addressed the issue for the single porosity equations, similar work remains to be done for integral- and multiple porosity-based models.

3.4.3.2. Four year research objectives

The four year research objectives are the following:

- investigate the properties of the analytical solutions of the Riemann problem for a continuous, anisotropic porosity field,
- extend the properties of such analytical solutions to discontinuous porosity fields,
- derive accurate and CPU-efficient approximate Riemann solvers for the solution of the conservation form of the porosity equations.

3.4.3.3. People

Vincent Guinot, Pascal Finaud-Guyot

3.4.3.4. External collaborations

Owing to the limited staff of the LEMON team, external collaborations will be sought with researchers in applied mathematics. Examples of researchers working in the field are

- Minh Le, Saint Venant laboratory, Chatou (France): numerical methods for shallow water flows, experience with the 2D, finite element/finte volume-based Telemac2D system.
- M.E. Vazquez-Cendon, Univ. Santiago da Compostela (Spain): finite volume methods for shallow water hydrodynamics and transport, developed Riemann solvers for the single porosity equations.
- A. Ferrari, R. Vacondio, S. Dazzi, P. Mignosa, Univ. Parma (Italy): applied mathematics, Riemann solvers for the single porosity equations.
- O. Delestre, Univ. Nice-Sophia Antipolis (France): development of numerical methods for shallow water flows (source term treatment, etc.)
- F. Benkhaldoun, Univ. Paris 13 (France): development of Riemann solvers for the porous shallow water equations.

.

3.4.4. Inland hydrobiological systems

3.4.4.1. State of the Art

Water bodies such as lakes or coastal lagoons (possibly connected to the sea) located in high human activity areas are subject to various kinds of stress such as industrial pollution, high water demand or bacterial blooms caused by freshwater over-enrichment. For obvious environmental reasons, these water resources have to be protected, hence the need to better understand and possibly control such fragile ecosystems to eventually develop decision-making tools. From a modelling point of view, they share a common feature in that they all involve interacting biological and hydrological processes. According to [33], models may be classified into two main types: “minimal dynamic models” and “complex dynamic models”. These two model types do not have the same objectives. While the former are more heuristic and rather depict the likelihood of considered processes, the latter are usually derived from fundamental laws of biochemistry or fluid dynamics. Of course, the latter necessitate much more computational resources than the former. In addition, controlling such complex systems (usually governed by PDEs) is by far more difficult that controlling the simpler ODE-driven command systems.
LEMON has already contributed both to the reduction of PDE models for the simulation of water confinement in coastal lagoons [34], [16] and to the improvement of ODE models in order to account for space-heterogeneity of bioremediation processes in water resources [14].

3.4.4.2. Four year research objectives

In collaboration with colleagues from the ANR-ANSWER project and colleagues from INRA, our ambition is to improve existing models of lagoon/marine ecosystems by integrating both accurate and numerically affordable coupled hydrobiological systems. A major challenge is to find an optimal trade-off between the level of detail in the description of the ecosystem and the level of complexity in terms of number of parameters (in particular regarding the governing equations for inter-species reactions). The model(s) should be able to reproduce the inter-annual variability of the observed dynamics of the ecosystem in response to meteorological forcing. This will require the adaptation of hydrodynamics equations to such time scales (reduced/upscaled models such as porosity shallow water models (see Section 3.2.1) will have to be considered) together with the coupling with the ecological models. At short time scales (i.e. the weekly time scale), accurate (but possibly CPU-consuming) 3D hydrodynamic models processes (describing thermal stratification, mixing, current velocity, sediment resuspension, wind waves...) are needed. On the longer term, it is intended to develop reduced models accounting for spatial heterogeneity.

The team will focus on two main application projects in the coming years:

- the ANR ANSWER project (2017-2021, with INRA Montpellier and LEESU) focusing on the cyanobacteria dynamics in lagoons and lakes. A PhD student is co-advised by Antoine Rousseau in collaboration with Céline Casenave (INRA, Montpellier).
- the long term collaboration with Alain Rapaport (INRA Montpellier) will continue both on the bioremediation of water resources such as the Tunquen lagoon in Chile and with a new ongoing project on water reuse (converting wastewater into water that can be reused for other purposes such as irrigation of agricultural fields). Several projects are submitted to the ANR and local funding structures in Montpellier.

3.4.4.3. People

Céline Casenave (INRA Montpellier), Antoine Rousseau, Vincent Guinot, Joseph Luis Kahn Casapia.

3.4.4.4. External collaborations

- ANR ANSWER consortium: Céline Casenave (UMR MISTEA, INRA Montpellier), Brigitte Vinçon-Leite (UM LEESU, ENPC), Jean-François Humbert (UMR IEES, UPMC). ANSWER is a French-Chinese collaborative project that focuses on the modelling and simulation of eutrophic lake ecosystems to study the impact of anthropogenic environmental changes on the proliferation of cyanobacteria. Worldwide the current environmental situation is preoccupying: man-driven water needs increase, while the quality of the available resources is deteriorating due to pollution of various kinds and to hydric stress. In particular, the eutrophication of lentic ecosystems due to excessive inputs of nutrients (phosphorus and nitrogen) has become a major problem because it promotes cyanobacteria blooms, which disrupt the functioning and the uses of the ecosystems.
- A. Rousseau has a long lasting collaboration with Alain Rapaport (UMR MISTEA, INRA Montpellier) and Héctor Ramirez (CMM, Université du Chili).

LIFEWARE Project-Team

3. Research Program

3.1. Computational Systems Biology

Bridging the gap between the complexity of biological systems and our capacity to model and **quantitatively predict system behaviors** is a central challenge in systems biology. We believe that a deeper understanding of the concept and theory of biochemical computation is necessary to tackle that challenge. Progress in the theory is necessary for scaling, and enabling the application of static analysis, module identification and decomposition, model reductions, parameter search, and model inference methods to large biochemical reaction systems. A measure of success on this route will be the production of better computational modeling tools for elucidating the complex dynamics of natural biological processes, designing synthetic biological circuits and biosensors, developing novel therapy strategies, and optimizing patient-tailored therapeutics.

Progress on the **coupling of models to data** is also necessary. Our approach based on quantitative temporal logics provides a powerful framework for formalizing experimental observations and using them as formal specification in model building. Key to success is a tight integration between *in vivo* and *in silico* work, and on the mixing of dry and wet experiments, enabled by novel biotechnologies. In particular, the use of microfluidic devices makes it possible to measure behaviors at both single-cell and cell population levels *in vivo*, provided innovative modeling, analysis and control methods are deployed *in silico*.

In synthetic biology, while the construction of simple intracellular circuits has shown feasible, the design of larger, **multicellular systems** is a major open issue. In engineered tissues for example, the behavior results from the subtle interplay between intracellular processes (signal transduction, gene expression) and intercellular processes (contact inhibition, gradient of diffusible molecule), and the question is how should cells be genetically modified such that the desired behavior robustly emerges from cell interactions.

3.2. Chemical Reaction Network (CRN) Theory

Feinberg's chemical reaction network theory and Thomas's influence network analyses provide sufficient and/or necessary structural conditions for the existence of multiple steady states and oscillations in regulatory networks. Those conditions can be verified by static analyzers without knowing kinetic parameter values nor making any simulation. In this domain, most of our work consists in analyzing the interplay between the **structure** (Petri net properties, influence graph, subgraph epimorphisms) and the **dynamics** (Boolean, CTMC, ODE, time scale separations) of biochemical reaction systems. In particular, our study of influence graphs of reaction systems, our generalization of Thomas' conditions of multi-stationarity and Soulé's proof to reaction systems [0], the inference of reaction systems from ODEs [0], the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms now provide solid ground for developing static analyzers, using them on a large scale in systems biology, and elucidating modules.

3.3. Logical Paradigm for Systems Biology

Our group was among the first ones in 2002 to apply **model-checking** methods to systems biology in order to reason on large molecular interaction networks, such as Kohn's map of the mammalian cell cycle (800 reactions over 500 molecules) [0]. The logical paradigm for systems biology that we have subsequently developed for quantitative models can be summarized by the following identifications :

[0]Sylvain Soliman. A stronger necessary condition for the multistationarity of chemical reaction networks. Bulletin of Mathematical Biology, 75(11):2289–2303, 2013.

[0]François Fages, Steven Gay, Sylvain Soliman. Inferring reaction systems from ordinary differential equations. Journal of Theoretical Computer Science (TCS), Elsevier, 2015, 599, pp.64–78.

[0]N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, V. Schächter. Modeling and querying biochemical interaction networks. Theoretical Computer Science, 325(1):25–44, 2004.

biological model = transition system K

dynamical behavior specification = temporal logic formula ϕ

model validation = model-checking $K,\ s \models ?\ \phi$

model reduction = sub-model-checking, $K' \subset K$ s.t. $K'?,\ s \models \phi$

model prediction = formula enumeration, ϕ s.t. $K,\ s \models \phi?$

static experiment design = symbolic model-checking, state s s.t. $K,\ s? \models \phi$

model synthesis = constraint solving $K?,\ s \models \phi$

dynamic experiment design = constraint solving $K?,\ s? \models \phi$

In particular, the definition of a continuous satisfaction degree for **first-order temporal logic** formulae with constraints over the reals, was the key to generalize this approach to quantitative models, opening up the field of model-checking to model optimization [0] This line of research continues with the development of temporal logic patterns with efficient constraint solvers and their generalization to handle stochastic effects.

3.4. Computer-Aided Design of CRNs for Synthetic Biology

The continuous nature of many protein interactions leads us to consider models of analog computation, and in particular, the recent results in the theory of analog computability and complexity obtained by Amaury Pouly [0] and Olivier Bournez, establish fundamental links with digital computation. In a paper published last year [0] We have derived from these results the Turing completeness result of elementary CRNs (without polymerization) under the differential semantics, closing a long-standing open problem in CRN theory. The proof of this result shows how computable function over the reals, described by Ordinary Differential Equations, namely by Polynomial Initial Value Problems (PIVP), can be compiled into elementary biochemical reactions, furthermore with a notion of analog computation complexity defined as the length of the trajectory to reach a given precision on the result. This opens a whole research avenue to analyze biochemical circuits in Systems Biology, transform behavioural specifications into biochemical reactions for Synthetic Biology, and compare artificial circuits with natural circuits acquired through evolution, from the novel point of view of analog computation and complexity.

3.5. Modeling of Phenotypic Heterogeneity in Cellular Processes

Since nearly two decades, a significant interest has grown for getting a quantitative understanding of the functioning of biological systems at the cellular level. Given their complexity, proposing a model accounting for the observed cell responses, or better, predicting novel behaviors, is now regarded as an essential step to validate a proposed mechanism in systems biology. Moreover, the constant improvement of stimulation and observation tools creates a strong push for the development of methods that provide predictions that are increasingly precise (single cell precision) and robust (complex stimulation profiles).

It is now fully apparent that cells do not respond identically to a same stimulation, even when they are all genetically-identical. This phenotypic heterogeneity plays a significant role in a number of problems ranging from cell resistance to anticancer drug treatments to stress adaptation and bet hedging.

[0] On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology A. Rizk, G. Batt, F. Fages, S. Soliman International Conference on Computational Methods in Systems Biology, 251-268

[0] Amaury Pouly, “Continuous models of computation: from computability to complexity”, PhD Thesis, Ecole Polytechnique, Nov. 2015.

[0] Fages, François, Le Guludec, Guillaume and Bournez, Olivier, Pouly, Amaury. Strong Turing Completeness of Continuous Chemical Reaction Networks and Compilation of Mixed Analog-Digital Programs. In CMSB’17: Proceedings of the fiveteen international conference on Computational Methods in Systems Biology, pages 108–127, volume 10545 of Lecture Notes in Computer Science. Springer-Verlag, 2017.

Dedicated modeling frameworks, notably **stochastic** modeling frameworks, such as chemical master equations, and **statistic** modeling frameworks, such as ensemble models, are then needed to capture biological variability.

Appropriate mathematical and computational tools should then be employed for the analysis of these models and their calibration to experimental data. One can notably mention **global optimization** tools to search for appropriate parameters within large spaces, **moment closure** approaches to efficiently approximate stochastic models [0], and (stochastic approximations of) the **expectation maximization** algorithm for the identification of mixed-effects models [0].

3.6. External Control of Cell Processes

External control has been employed since many years to regulate culture growth and other physiological properties. Recently, taking inspiration from developments in synthetic biology, closed loop control has been applied to the regulation of intracellular processes. Such approaches offer unprecedented opportunities to investigate how a cell process dynamical information by maintaining it around specific operating points or driving it out of its standard operating conditions. They can also be used to complement and help the development of synthetic biology through the creation of hybrid systems resulting from the interconnection of in vivo and in silico computing devices.

In collaboration with Pascal Hersen (CNRS MSC lab), we developed a platform for gene expression control that enables to control protein concentrations in yeast cells. This platform integrates microfluidic devices enabling long-term observation and rapid change of the cells environment, microscopy for single cell measurements, and software for real-time signal quantification and model based control. We demonstrated in 2012 that this platform enables controlling the level of a fluorescent protein in cells with unprecedented accuracy and for many cell generations [0].

More recently, motivated by an analogy with a benchmark control problem, the stabilization of an inverted pendulum, we investigated the possibility to balance a genetic toggle switch in the vicinity of its unstable equilibrium configuration. We searched for solutions to balance an individual cell and even an entire population of heterogeneous cells, each harboring a toggle switch [0].

Independently, in collaboration with colleagues from IST Austria, we investigated the problem of controlling cells, one at a time, by constructing an integrated optogenetic-enabled microscopy platform. It enables experiments that bridge individual and population behaviors. We demonstrated: (i) population structuring by independent closed-loop control of gene expression in many individual cells, (ii) cell–cell variation control during antibiotic perturbation, (iii) hybrid bio-digital circuits in single cells, and freely specifiable digital communication between individual bacteria [0].

3.7. Constraint Solving and Optimization

Constraint solving and optimization methods are important in our research. On the one hand, static analysis of biochemical reaction networks involves solving hard combinatorial optimization problems, for which **constraint programming** techniques have shown particularly successful, often beating dedicated algorithms

[0]Moment-based inference predicts bimodality in transient gene expression, C. Zechner C, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, Proceedings of the National Academy of Sciences USA, 9(5):109(21):8340-5, 2012

[0]What population reveals about individual cell identity: estimation of single-cell models of gene expression in yeast, A. Llamosi, A.M. Gonzalez-Vargas, C. Versari, E. Cinquemani, G. Ferrari-Trecate, P. Hersen, and G. Batt, PLoS Computational Biology, 9(5): e1003056, 2015

[0]Jannis Uhlendorf, Agnés Miermont, Thierry Delaveau, Gilles Charvin, François Fages, Samuel Bottani, Grégory Batt, Pascal Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. Proceedings of the National Academy of Sciences USA, 109(35):14271–14276, 2012.

[0]Jean-Baptiste Lugagne, Sebastian Sosa Carrillo and Melanie Kirch, Agnes Köhler, Gregory Batt and Pascal Hersen. Balancing a genetic toggle switch by real-time feedback control and periodic forcing. Nature Communications, 8(1):1671, 2017.

[0]Remy Chait, Jakob Ruess, Tobias Bergmiller and Gavsper Tkavcik, Cvalin Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. Nature Communications, 8(1):1535, 2017.

and allowing to solve large instances from model repositories. On the other hand, parameter search and model calibration problems involve similarly solving hard continuous optimization problems, for which **evolutionary algorithms**, and especially the covariance matrix evolution strategy (CMA-ES) [0] have been shown to provide best results in our context, for up to 100 parameters. This has been instrumental in building challenging quantitative models, gaining model-based insights, revisiting admitted assumptions, and contributing to biological knowledge [00].

[0]N. Hansen, A. Ostermeier (2001). Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 9(2) pp. 159–195.

[0]Domitille Heitzler, Guillaume Durand, Nathalie Gallay, Aurélien Rizk, Seungkirl Ahn, Jihee Kim, Jonathan D. Violin, Laurence Dupuy, Christophe Gauthier, Vincent Piketty, Pascale Crépieux, Anne Poupon, Frédérique Clément, François Fages, Robert J. Lefkowitz, Eric Reiter. Competing G protein-coupled receptor kinases balance G protein and β-arrestin signaling. Molecular Systems Biology, 8(590), 2012.

[0]Pauline Traynard, Céline Feillet, Sylvain Soliman, Franck Delaunay, François Fages. Model-based Investigation of the Circadian Clock and Cell Cycle Coupling in Mouse Embryonic Fibroblasts: Prediction of RevErb-alpha Up-Regulation during Mitosis. Biosystems, 149:59–69, 2016.

M3DISIM Project-Team

3. Research Program

3.1. Multi-scale modeling and coupling mechanisms for biomechanical systems, with mathematical and numerical analysis

Over the past decade, we have laid out the foundations of a multi-scale 3D model of the cardiac mechanical contraction responding to electrical activation. Several collaborations have been crucial in this enterprise, see below references. By integrating this formulation with adapted numerical methods, we are now able to represent the whole organ behavior in interaction with the blood during complete heart beats. This subject was our first achievement to combine a deep understanding of the underlying physics and physiology and our constant concern of proposing well-posed mathematical formulations and adequate numerical discretizations. In fact, we have shown that our model satisfies the essential thermo-mechanical laws, and in particular the energy balance, and proposed compatible numerical schemes that – in consequence – can be rigorously analyzed, see [6]. In the same spirit, we have formulated a poromechanical model adapted to the blood perfusion in the heart, hence precisely taking into account the large deformation of the mechanical medium, the fluid inertia and moving domain, and so that the energy balance between fluid and solid is fulfilled from the model construction to its discretization, see [7].

3.2. Inverse problems with actual data – Fundamental formulation, mathematical analysis and applications

A major challenge in the context of biomechanical modeling – and more generally in modeling for life sciences – lies in using the large amount of data available on the system to circumvent the lack of absolute modeling ground truth, since every system considered is in fact patient-specific, with possibly non-standard conditions associated with a disease. We have already developed original strategies for solving this particular type of inverse problems by adopting the observer stand-point. The idea we proposed consists in incorporating to the classical discretization of the mechanical system an estimator filter that can use the data to improve the quality of the global approximation, and concurrently identify some uncertain parameters possibly related to a diseased state of the patient. Therefore, our strategy leads to a coupled model-data system solved similarly to a usual PDE-based model, with a computational cost directly comparable to classical Galerkin approximations. We have already worked on the formulation, the mathematical and numerical analysis of the resulting system – see [5] – and the demonstration of the capabilities of this approach in the context of identification of constitutive parameters for a heart model with real data, including medical imaging, see [3].

MAGIQUE-3D Project-Team

3. Research Program

3.1. Introduction

Probing the invisible is a quest that is shared by a wide variety of scientists such as archaeologists, geologists, astrophysicists, physicists, etc... Magique-3D is mainly involved in Geophysical imaging which aims at understanding the internal structure of the Earth from the propagation of waves. Both qualitative and quantitative information are required and two geophysical techniques can be used: **seismic reflection** and **seismic inversion**. Seismic reflection provides a qualitative description of the subsurface from reflected seismic waves by indicating the position of the reflectors while seismic inversion transforms seismic reflection data into a quantitative description of the subsurface. Both techniques are inverse problems based upon the numerical solution of wave equations. Oil and Gas explorations have been pioneering application domains for seismic reflection and inversion and even if numerical seismic imaging is computationally intensive, oil companies clearly promote the use of numerical simulations to provide synthetic maps of the subsurface. This is due to the tremendous progresses of scientific computing which have pushed the limits of existing numerical methods and it is now conceivable to tackle realistic 3D problems. However, mathematical wave modeling has to be well-adapted to the region of interest and the numerical schemes which are employed to solve wave equations have to be both accurate and scalable enough to take full advantage of parallel computing. Today, geophysical imaging tackles more and more realistic problems and we contribute to this task by improving the modeling and by deriving advanced numerical methods for solving wave problems.
MAGIQUE-3D research program is divided into four axes that are: (1) Imaging the Earth; (2) Exploring the Sun; (3) Detecting defaults in complex media; (4) Designing objects with a variety of shapes. Those applications stand out from the collaborations that we have established with interested end-user groups. It is worth noting that they share basic common methodologies which imparts consistency to our program despite they may appear quite distant. MAGIQUE-3D keep modeling and simulating geophysical phenomena for understanding the Earth interior and developing its resources sustainably, and our xperience with numerical geophysics may help us to address other challenging applications. We mainly used DG finite elements and spectral elements and both have demonstrated very good performance. However, in particular for reducing the computational costs and/or for better capturing the propagation characteristics, we are working on the development of hybrid solvers based on the coupling of different finite element methods. Other open questions deserve attention like the problem of numerical pollution or the poor scalability of decomposition domain techniques which are both significantly hampering computations in very large domains. For those purposes, we focus ourselves on the development of Trefftz-like approximations that are based on a particular computation of the fluxes by making a judicious use of an auxiliary numerical method (e.g. boundary integral equations, spectral elements, etc...). Those problems cannot be ignored and can be found in all of our research axes. In addressing those issues, we participate in the construction of new numerical schemes and for that purpose, we continuously need to improve our understanding of the underlying physics. By this way, we make our mathematical models evolve to more realistic representations of the wave propagation phenomenon. This motivated us to introduce experimental studies in our activities and to collaborate with geophysicists of the UPPA who own experimental devices adapted to our concerns. Moreover, we have hired recently Yder Masson who is an experienced researcher developing modeling and imaging methods to investigate the Earth's internal structure. This creates all the conditions for improving our mathematical representation of waves in complex media. It is worth noting that modeling is a concern for both geophysicists and mathematicians. Indeed, the Physics must be reproduced accurately and the underlying mathematical properties should be clarified. By this way, we can develop a numerical scheme respecting the main properties of the continuous problem of interest (energy conservation or attenuation, stability, well-posedness, etc...). Magique-3D proposes to define its research program from in-house accurate solution methodologies for simulating wave propagation in realistic scenarios to various applications involving trans-disciplinary efforts. The development of high-order

numerical methods for wave simulations is serving as a basis for our contributions regarding applications. In particular, we pursue and strengthen our collaboration with HPC teams, in order to improve the scalability of our codes and to run them on very large heterogeneous architectures (using task based programming libraries as StarPU developed by Inria project-team Storm, improving the I/O by collaborating with UFRGS at Porto Alegre, using the metaprogramming framework Boast developed by Inria project-team Corse to produce portable and efficient computing kernels). We are also continuing our collaboration with Inria project team Hiepacs on the use of hybrid linear solvers, by considering the multiple Right-Hand Sides feature and by integrating appropriate transmission conditions between the various domains. During 2019, we have worked a lot on: (a) High-order numerical methods for modeling wave propagation in porous media: development and implementation; (b) Understanding the interior of the Earth and the Sun by solving inverse problems; (c) Full waveform inversion for the optimal design of wind musical instruments.

3.2. High-order numerical methods for modeling wave propagation in porous media: development and implementation

We aim at achieving the characterization of conducting porous media which are media favoring the conversion of seismic waves into electromagnetic waves.. This project is identified as a ”New scientific challenge” which is a set of research projects funded by the E2S project of UPPA. The shape and form of porous media can vary depending on the size of the pore and the structure of the solid skeleton. Porous media are found in the nature (sandstone, volcanic rocks, etc) or can be manufactured (concrete, polyurethane foam, etc) as depicted in [68]. Instead of modeling such media as strongly heterogeneous, homogenization is used to describe the material on a macroscopic scale. Biot's theory describes the solid skeleton according to linear elasticity and adds to this the Navier-Stokes equation for a viscous fluid and Darcy's law governing the motion of the fluid [63], [61]. For simplified linear elasticity, there are one equation of motion and one constitutive law, with the unknowns being the displacement field in the solid and the solid stress. In poro-elasticity, the added unknowns are the fluid displacement relative to the solid and the fluid pressure. There are two equations of motion, coupled with two constitutive laws. By plane wave analysis, one obtains three types of waves: S wave, fast P wave and slow P wave (Biot's wave). While the first two types are similar to those existing in elastic solid, the existence of a third wave with drastically smaller speed adds to the complications already encountered in elasticity. This is obviously even more challenging for conducting poroelastic media where the three poroelastic waves are coupled with an electric field. In this case, it is not realistic to use a unique scheme for all the waves. Standard finite element methods coupled with time schemes have indeed difficulties to deliver accurate solutions because there is a need of adapting the mesh size to the smallest wave velocity and the time discretization to the largest wave velocity. It is then tricky to numerically reproduce the Biot's wave while approximating correctly the regular elastic waves P and S. Moreover, there is a challenging question about the boundary condition to be used for limiting the computational domain. We have launched a Ph.D project (Rose-Cloé Meyer) aiming at developing a new piece of software for the simulation of time-harmonic waves in conducting porous media. This project is developed in collaboration with Steve Pride from the Lawrence Berkeley National Laboratory who has elaborated the corresponding physical theory [83], [87], [73], [74]. Next, once a new numerical method is developed, it is validated by comparing the numerical solution to an analytical one. This is a key step to us for assessing the accuracy of our simulations. Nevertheless, analytical solutions are not available for realistic media such as poroelastic or viscoelastic media represented by heterogeneous parameters. Engineers still argue that simulations may be inaccurate and could lead to wrong conclusions. Fortunately, it is possible to produce experimentally quite complex configurations where multi-physics measurements are used to monitor the wave propagation. There is thus a possibility of moving further on the validation of the numerical methods by comparing simulations and experiments. What is very exciting is that experiments are used to validate numerical methods which have the objective of simulating new phenomena that are not possible to reproduce in a lab. We have launched two Ph.D thesis (Chengyi Shen and Victor Martins Gomes) in collaboration with Daniel Brito (LFCR-UPPA) on the comparison of simulations with experiments. This topic is connected to another project that we have with Total on the use of waves for characterizing carbonates.

3.3. Understanding the interior of the Earth and the Sun by solving inverse problems

Even if the Earth and the Sun are actually very different media, their imaging is based on the same solution methodology [64]. However, our knowledge on Earth inversion is far more developed than for the Sun. Earth inversion is in the continuation of previous MAGIQUE-3D achievements while Sun inversion requires developing new technologies based on modeling, numerical analysis and implementation of a piece of software which is able to ask for new developments. For instance, we would like to develop a HDG software package for solving Galbrun and Linearized Euler equations. To the best of our knowledge, this has never be done and would be a major milestone for tackling vectorial equations. Regarding the modeling, we are pursuing our collaboration with the Max Planck Institute for Solar System Research (Göttingen, Germany) in the framework of the associate team ANTS. This partnership is essential to us for understanding a complex (and new to us) physics including gravity waves that we have never considered in the past. Even if we dispose of advanced solvers dealing with elasticity, the development of fast and accurate solvers for reproducing waves travelling in large 3D domains is still one of the positive developments towards realistic simulations. In particular, the techniques for the forward discretization and linear system solver must evolve accordingly to resolve large scale time-harmonic problems. For instance, we have elaborated a space-time Trefftz-DG formulation of the elasto-acoustic problem [58], which performs very well regarding the number of dofs and the order of convergence. We have also coupled spectral and DG elements to take advantage of both methods and we have performed some simulations which are very promising [57]. The formulation of FWI is in progress in the framework of Pierre Jacquet thesis launched in November 2017. Finally, we have also initiated research on seismology at the planetary scale, with the arrival of Yder Masson on the subject and new collaborators (such as Berkeley lab). This will further help widen our expertise on inverse wave problems and will feed all the four research axes of the future team-project. Regarding industrial partnerships, we have collaboration with Total and the SME RealtimeSeismic (Pau, France). We also continue to work with the UPV, the BCAM and the BSC, namely in the framework of Mathrocks project.

3.4. Hybrid time discretizations of high-order

Most of the meshes we consider are composed of cells greatly varying in size. This can be due to the physical characteristics (propagation speed, topography, ...) which may require to refine the mesh locally, very unstructured meshes can also be the result of dysfunction of the mesher. For practical reasons which are essentially guided by the aim of reducing the number of matrix inversions, explicit schemes are generally privileged. However, they work under a stability condition, the so-called Courant Friedrichs Lewy (CFL) condition which forces the time step being proportional to the size of the smallest cell. Then, it is necessary to perform a huge number of iterations in time and in most of the cases because of a very few number of small cells. This implies to apply a very small time step on grids mainly composed of coarse cells and thus, there is a risk of creating numerical dispersion that should not exist. However, this drawback can be avoided by using low degree polynomial basis in space in the small meshes and high degree polynomials in the coarse meshes. By this way, it is possible to relax the CFL condition and in the same time, the dispersion effects are limited. Unfortunately, the cell-size variations are so important that this strategy is not sufficient. One solution could be to apply implicit and unconditionally stable schemes, which would obviously free us from the CFL constraint. Unfortunately, these schemes require inverting a linear system at each iteration and thus needs huge computational burden that can be prohibitive in 3D. Moreover, numerical dispersion may be increased. Then, as second solution is the use of local time stepping strategies for matching the time step to the different sizes of the mesh. There are several attempts [65], [60], [82], [79], [71] and Magique 3D has proposed a new time stepping method which allows us to adapt both the time step and the order of time approximation to the size of the cells. Nevertheless, despite a very good performance assessment in academic configurations, we have observed to our detriment that its implementation inside industrial codes is not obvious and in practice, improvements of the computational costs are disappointing, especially in a HPC framework. Indeed, the local time stepping algorithm may strongly affect the scalability of the code. Moreover, the complexity of the algorithm is increased when dealing with lossy media [76].

Recently, Dolean *et al* [70] have considered a novel approach consisting in applying hybrid schemes combining second order implicit schemes in the thin cells and second order explicit discretization in the coarse mesh. Their numerical results indicate that this method could be a good alternative but the numerical dispersion is still present. It would then be interesting to implement this idea with high-order time schemes to reduce the numerical dispersion. The recent arrival in the team of J. Chabassier should help us to address this problem since she has the expertise in constructing high-order implicit time scheme based on energy preserving Newmark schemes [62]. We propose that our work be organized around the two following tasks. The first one is the extension of these schemes to the case of lossy media because applying existing schemes when there is attenuation is not straightforward. This is a key issue because there is artificial attenuation when absorbing boundary conditions are introduced and if not, there are cases with natural attenuation like in visco-elastic media. The second one is the coupling of high-order implicit schemes with high-order explicit schemes. These two tasks can be first completed independently, but the ultimate goal is obviously to couple the schemes for lossy media. We will consider two strategies for the coupling. The first one will be based on the method proposed by Dolean *et al*, the second one will consist in using Lagrange multiplier on the interface between the coarse and fine grids and write a novel coupling condition that ensures the high order consistency of the global scheme. Besides these theoretical aspects, we will have to implement the method in industrial codes and our discretization methodology is very suitable for parallel computing since it involves Lagrange multipliers. We propose to organize this task as follows. There is first the crucial issue of a systematic distribution of the cells in the coarse/explicit and in the fine/implicit part. Based on our experience on local time stepping, we claim that it is necessary to define a criterion which discriminates thin cells from coarse ones. Indeed, we intend to develop codes which will be used by practitioners, in particular engineers working in the production department of Total. It implies that the code will be used by people who are not necessarily experts in scientific computing. Considering real-world problems means that the mesh will most probably be composed of a more or less high number of subsets arbitrarily distributed and containing thin or coarse cells. Moreover, in the prospect of solving inverse problems, it is difficult to assess which cells are thin or not in a mesh which varies at each iteration.

Another important issue is the load balancing that we can not avoid with parallel computing. In particular, we will have to choose one of these two alternatives: dedicate one part of processors to the implicit computations and the other one to explicit calculus or distribute the resolution with both schemes on all processors. A collaboration with experts in HPC is then mandatory since we are not expert in parallel computing. We will thus continue to collaborate with the team-projects Hiepacs and Runtime with whom we have a long-term experience of collaborations. The load-balancing leads then to the issue of mesh partitioning. Main mesh partitioners are very efficient for the coupling of different discretizations in space but to the best of our knowledge, the case of non-uniform time discretization has never been addressed. The study of meshes being out of the scopes of Magique-3D, we will collaborate with experts on mesh partitioning. We get already on to François Pellegrini who is the principal investigator of Scotch (http://www.labri.fr/perso/pelegrin/scotch) and permanent member of the team project Bacchus (Inria Bordeaux Sud Ouest Research Center).

In the future, we aim at enlarging the application range of implicit schemes. The idea will be to use the degrees of freedom offered by the implicit discretization in order to tackle specific difficulties that may appear in some systems. For instance, in systems involving several waves (as P and S waves in porous elastic media, or coupled wave problems as previously mentioned) the implicit parameter could be adapted to each wave and optimized in order to reduce the computational cost. More generally, we aim at reducing numeric bottlenecks by adapting the implicit discretization to specific cases.

3.5. Full waveform inversion for the optimal design of wind musical instruments

Makers have improved wind musical instruments (as flutes, trumpets, clarinets, bassoons, ...) in the past by a “trial and error” procedure, where the final sound and ease of the instrument in playing conditions are the main criteria. Although the playing context should still be the final reference, we can consider intermediate measurements of the pipe entry impedance [75], [69], which quantifies the Dirichlet-to-Neumann map of the

wave propagation in the pipe, and relies on mathematical simulations based on accurate and concise models of the pipe [84], [59] and the embouchure [77], [59], [55], [56], [86] in order to foresee the behavior of a given instrument, and therefore optimize it. A strong interaction with makers and players is necessary for defining both operable criteria quantified as a cost function and a design parameters space. We aim at building efficient musical instrument via handcrafted techniques but also modern tools as additive synthesis (3D printers). We plan to implement state-of-the-art numerical methods (finite elements, full waveform inversion, neuronal networks fed by numerical simulations, diverse optimization techniques...) that are versatile (in terms of models, formulations, couplings...) in order to solve the optimization problem, after a proper modeling of the linear and nonlinear coupled phenomena. We wish to take advantage of the fact that sound waves in musical instruments satisfy the laws of acoustics in pipes (PDEs), which leads to use FWI technique, in harmonic or temporal regime. We propose to implement an iterative process between instrument making and optimal design in order to build instruments that optimize tone quality and playability. We are currently collaborating with musical acoustics teams who have a strong experimental background on this question [66], [67] [DCY12, DF07, GPP98], we wish to strenghten the links we have with other teams [78], [72], [80], [81], [85], we will participate to professional clusters [ITE], and we are currently collaborating with makers and museums directly : Augustin Humeau (Dordogne) for the bassoon, Luc Gallois (Oise) and the Museum of Cité de la Musique - Philharmonie de Paris for the brass instruments. This research axis is surely the most exploratory of our research program and follows the successful ”Exploratory Research Program” Inria grant obtained in 2017. It could pave the way for significant progresses in inverse problem solving. Indeed, the problem depends on a few number of parameters unlike geophysical or astrophysical problems. We can thus use it to test different methods like neuronal networks, statistical methods, coupling with nonlinear phenomena, and decide if it could be applied to large scale applications.

MAMBA Project-Team

3. Research Program

3.1. Introduction

Data and image analysis, statistical, ODEs, PDEs, and agent-based approaches are used either individually or in combination, with a strong focus on PDE analysis and agent-based approaches. Mamba was created in January 2014. It aims at developing models, simulations, numerical and control algorithms to solve questions from life sciences involving dynamics of phenomena encountered in biological systems such as protein intracellular spatio-temporal dynamics, cell motion, early embryonic development, multicellular growth, wound healing and liver regeneration, cancer evolution, healthy and tumor growth control by pharmaceuticals, protein polymerization occurring in neurodegenerative disorders, control of dengue epidemics, etc.

Another guideline of our project is to remain close to the most recent questions of experimental biology or medicine.In this context, we develop many close and fruitful collaborations with biologists and physicians, among which we can quote: the collaboration with St Antoine Hospital in Paris within the Institut Universitaire de Cancérologie of Sorbonne Université (IUC, Luis Almeida, Jean Clairambault, Dirk Drasdo, Benoît Perthame); Institut Jacques Monod (Luis Almeida); INRA Jouy-en-Josas (VIM team, headed by Human Rezaei and Vincent Béringue (Marie Doumic anc Philippe Robert); Wei-Feng Xue's team in the university of Canterbury (Marie Doumic and Philippe Robert); our collaborators within the HTE program (François Delhommeau at St Antoine, Thierry Jaffredo, and Delphine Salort at IBPS, Sorbonne Université, Paris; François Vallette at INSERM Nantes); Frédéric Thomas at CREEC, Montpellier; Hôpital Paul Brousse through ANR-IFlow and ANR-iLite; Institut de Biologie Physico-Chimique (IBPC, Paris, Teresa Teixeira's team; Marie Doumic); the close experimental collaborations that emerged through the former associated team QUANTISS (Dirk Drasdo), particularly at the Leibniz Institute for Working Environment and Human Factors in Dortmund, Germany; Yves Dumont at CIRAD, Montpellier.

We focus mainly on the creation, investigation and transfer of new mathematical models, methods of analysis and control, and numerical algorithms, but in selected cases software development as that of CellSys and TiQuant by D. Drasdo and S. Hoehme is performed. More frequently, the team develops "proof of concept" numerical codes in order to test the adequacy of our models to experimental biology.

We have organized the presentation of our research program in three methodological axes (Subsections 3.2 , 3.3 and 3.4) and two application axes (Subsections 4.2 and 4.3). Evolving along their own logic in close interaction with the methodological axes, the application axes are considered as application-driven research axes in themselves. The methodological research axes are the following.

Axis 1 is devoted to work in physiologically-based design, analysis and control of population dynamics. It encompasses populations of bacteria, of yeasts, of cancer cells, of neurons, of aggregating proteins, etc. whose dynamics are represented by partial differential equations (PDEs), structured in evolving physiological traits, such as age, size, size-increment, time elapsed since last firing (neurons).

Axis 2 is devoted to reaction equations and motion equations of agents in living systems. It aims at describing biological phenomena such as tumor growth, chemotaxis and wound healing.

Axis 3 tackles the question of model and parameter identification, combining stochastic and deterministic approaches and inverse problem methods in nonlocal and multi-scale models.

3.2. Methodological axis 1: analysis and control for population dynamics

Personnel Pierre-Alexandre Bliman, Jean Clairambault, Marie Doumic, Benoît Perthame, Diane Peurichard, Nastassia Pouradier Duteil, Philippe Robert

Project-team positioning

Population dynamics is a field with varied and wide applications, many of them being in the core of MAMBA interests - cancer, bacterial growth, protein aggregation. Their theoretical study also brings a qualitative understanding on the interplay between individual growth, propagation and reproduction in such populations. In the past decades, many results where obtained in the BANG team on the asymptotic and qualitative behavior of such structured population equations, see e.g. [135], [73], [99], [84]. Other Inria teams interested by this domain are Mycenae, Numed and Dracula, with which we are in close contacts. Among the leaders of the domain abroad, we can cite among others our colleagues Tom Banks (USA), Graeme Wake (New Zealand), Glenn Webb (USA), Jacek Banasiak (South Africa), Odo Diekmann (Netherlands), with whom we are also in regular contact. Most remarkably and recently, connections have also been made with probabilists working on Piecewise Deterministic Markov Processes (F. Malrieu at the university of Rennes, Jean Bertoin at the ETH in Zurich, Vincent Bansaye at Ecole Polytechnique, Julien Berestycki at Cambridge, Amaury Lambert at College de France, M. Hoffmann at Paris Dauphine, Alex Watson in UCL, London and J. Bertoin in Zurich), leading to a better understanding of the links between both types of results – see also the Methodological axis 3.

Scientific achievements

We divide this research axis, which relies on the study of structured population equations, according to four different applications, bringing their own mathematical questions, e.g., stability, control, or blow-up.

Time asymptotics for nucleation, growth and division equations

Following the many results obtained in the BANG team on the asymptotic and qualitative behavior of structured population equation, we put our effort on the investigation of limit cases, where the trend to a steady state or to a steady exponential growth described by the first eigenvector fails to happen. In [78], the case of equal mitosis (division into two equally-sized offspring) with linear growth rate was studied, and strangely enough, it appeared that the general relative entropy method could also be adapted to such a non-dissipative case. Many discussions and common workshops with probabilists, especially through the ANR project PIECE coordinated by F. Malrieu, have led both communities to work closer.

In [96], the case of constant fragmentation rate and linear growth rate has been investigated in a deterministic approach, whereas similar questions were simultaneously raised but in a stochastic process approach in [75].

We also enriched the models by taking into account a nucleation term, modeling the spontaneous formation of large polymers out of monomers [147]. We investigated the interplay between four processes: nucleation, polymerization, depolymerization and fragmentation.

New perspectives are now to consider not only one species but several interacting ones, which may exhibit complex interplays which may lead to damped oscillations or to infinite growth; these are in collaboration with C. Schmeiser and within the Vienna associated team MaMoCeMa (J. Delacour's Ph.D) and with K. Fellner from Graz (M. Mezache's Ph.D).

Cell population dynamics and its control

One of the important incentives for such model design, source of many theoretical works, is the challenging question of drug-induced drug resistance in cancer cell populations, described in more detail below in the Applicative axis 1, Cancer. The adaptive dynamics setting used consists of phenotype-structured integro-differential [or reaction-diffusion, when phenotype instability is added under the form of a Laplacian] equations describing the dynamic behavior of different cell populations interacting in a Lotka-Volterra-like manner that represents common growth limitation due to scarcity of expansion space and nutrients. The phenotype structure allows us to analyse the evolution in phenotypic traits of the populations under study and its asymptotics for two populations [128], [125], [124], [126]. Space may be added as a complementary structure variable provided that something is known of the (Cartesian) geometry of the population [127], which is seldom the case.

Modelling Mendelian and non-Mendelian inheritances in density-dependent population dynamics

Classical strategies for controlling mosquitoes responsible of vector-borne disease are based on mechanical methods, such as elimination of oviposition sites; and chemical methods, such as insecticide spraying. Long term usage of the latter generates resistance [81], [110], transmitted to progeny according to Mendelian inheritance (in which each parent contributes randomly one of two possible alleles for a trait). New control strategies involve biological methods such as genetic control, which may either reduces mosquito population in a specific area or decreases the mosquito vector competence [61], [120], [156]. Among the latter, infection of wild populations by the bacterium *Wolbachia* appears promising (see also Applicative axis 2 below). Being maternally-transmitted, the latter obeys non-Mendelian inheritance law. Motivated by the effects of the (possibly unwanted) interaction of these two types of treatment, we initiated the study of modelling of Mendelian and non-Mendelian inheritances in density-dependent population dynamics. First results are shown in [59].

Control of collective dynamics

The term *self-organization* is used to describe the emergence of complex organizational patterns from simple interaction rules in collective dynamics systems. Such systems are valuable tools to model various biological systems or opinion dynamics, whether it be the collective movement of animal groups, the organization of cells in an organism or the evolution of opinions in a large crowd. A special case of self-organization is given by *consensus*, i.e. the situation in which all agents' state variables converge. Another phenomenon is that of *clustering*, when the group is split into clusters that each converge to a different state. We have designed optimal control strategies to drive collective dynamics to consensus. In the case where consensus and clustering are situations to be avoided (for example in crowd dynamics), we designed control strategies to keep the system away from clustering.

Models of neural network

Mean field limits have been proposed by biophysicists in order to describe neural networks based on physiological models. The various resulting equations are called integrate-and-fire, time elapsed models, voltage-conductance models. Their specific nonlinearities and the blow-up phenomena make their originality which has led to develop specific mathematical analysis [138], followed by [134], [119], [139], [83]. This field also yields a beautiful illustration for the capacity of the team to combine and compare stochastic and PDE modelling (see Methodological axis 3), in [89].

Models of interacting particle systems

The organisation of biological tissues during development is accompanied by the formation of sharp borders between distinct cell populations. The maintenance of this cell segregation is key in adult tissue homeostatis, and its disruption can lead tumor cells to spread and form metastasis. This segregation is challenged during tissue growth and morphogenesis due to the high mobility of many cells that can lead to intermingling. Therefore, understanding the mechanisms involved in the generation and maintain of cell segregation is of tremendous importance in tissue morphogenesis, homeostasis, and in the development of various invasive diseases such as tumors. In this research axis, we aim to provide a mathematical framework which enables to quantitatively link the segregation and border sharpening ability of the tissue to these cell-cell interaction phenomena of interest [72]. As agent-based models do not enable precise mathematical analysis of their solutions due to the lack of theoretical results, we turn towards continuous -macroscopic- models and aim to provide a rigorous link between the different models [71].

Models of population dynamics structured in phenotype

The collaboration of Jean Clairambault with Emmanuel Trélat and Camille Pouchol (from September this year assistant professor at MAP5 Paris-Descartes, University of Paris), together now with Nastassia Pouradier Duteil, has been continued and presently leads us to a possible quantitative biological identification of the structuring phenotypes of the model developed in [146], through a beginning collaboration with an Indian systems biologist (Mohit Kumar Jolly, IIS Bangalore). Our motivation in this collaboration is to couple a physiologically based system of 6 ODEs developed by our Indian collaborator with our phenotype-structured cell population dynamics model [13], [45].

In the framework of the HTE project EcoAML 2016-2020, Thanh Nam Nguyen, Jean Clairambault, Delphine Salort and Benoît Perthame, in collaboration with Thierry Jaffredo at IBPS-SU, have designed a phenotype-structured integrodifferential model of interactions between haematopoietic stem cells (healthy or leukaemic) and their supporting stromal cells [24]. In this model, without diffusion, to our relative astonishment, our postdoctoral fellow T.N. Nguyen predicts in particular that under special circumstances, a coexistence between healthy and leukaemic stem cell subpopulations is possible.The explanation of such possible theoretical coexistence still remains to be explained.

The idea of cooperation between cell subpopulations in a tumour is also studied using phenoytype-structured models of cell populations by Frank Ernesto Alvarez Borges, PhD student of Stéphane Mischler (Paris-Dauphine University), Mariano Rodríguez Ricard (University of Havana, Cuba) and Jean Clairambault, in collaboration with José Antonio Carrillo (Imperial College London). A feature of these models, in as much as conflicting continuous phenotypes (e.g., adhesivity vs. motility, or fecundity vs. viability, or fecundity vs. motility [0]) are supposed to structure a unique cell population, is that they can also represent the emergence of multicellularity in such a cell population, when two subpopulations of the same population, i.e., endowed with the same genome and represented w.r.t. relevant heterogeneity in the cell population by such conflicting phenotypes, are determined by two different choices of the 2-d phenotype. In a simplified representation when the two phenotypes are just extreme values of a 1-d continuous phenotype (e.g., 0 for total adhesivity and no motility, 1 for no adhesivity and complete motility) this situation may be related to the previously described case, developed in [24], in which two extreme values of a convex function linked to proliferation are occupied by the two extreme phenotype values (0 and 1), leading to the coexistence of two cell subpopulations.

Collaborations

- Nucleation, growth and fragmentation equations: **Klemens Fellner**, university of Graz, Austria, **Piotr Gwiązda**, Polish Academy of Sciences, Poland, **Christian Schmeiser**, university of Vienna.
- Cell population dynamics and its control: **Tommaso Lorenzi**, former Mamba postdoc, now at the University of St. Andrews, Scotland, maintains a vivid collaboration with the Mamba team. He is in particular an external member of the HTE program MoGlImaging (see also Applicative axis 1). **Emmanuel Trélat**, Sorbonne Université professor, member of LJLL and of the CAGE Inria team, is the closest Mamba collaborator for optimal control. **Benedetto Piccoli**, Professor at Rutgers University (Camden, New Jersey), is collaborating on the analysis and control of collective dynamics.
- Mendelian inheritance and resistance in density-dependent population dynamics: **Pastor Pérez-Estigarribia, Christian Schaerer**, Universidad Nacional de Asunción, Paraguay.
- Neural networks: **Delphine Salort,** Professor Sorbonne Université, Laboratory for computations and quantification in biology, and **Patricia Reynaud**, University of Nice, **Maria Cáceres**, University of Granada.
- Models of interacting particle systems: **Pierre Degond**, Imperial College London, Julien Barré, **MAPMO, Orléans**, **Ewelina Zatorska**, University College London

3.3. Methodological axis 2: reaction and motion equations for living systems

Personnel

Luis Almeida, Benoît Perthame, Diane Peurichard, Nastassia Pouradier Duteil, Dirk Drasdo

Project-team positioning

The Mamba team had initiated and is a leader on the works developed in this research axis. It is a part of a consortium of several mathematicians in France through the ANR Blanc project *Kibord*, which involves in particular members from others Inria team (DRACULA, COMMEDIA). Finally, we mention that from Sept. 2017 on, Mamba benefited from the ERC Advanced Grant ADORA (Asymptotic approach to spatial and dynamical organizations) of Benoît Perthame.

[0] as proposed by John Maynard Keynes and Eös Száthmary in their book "The major transitions in evolution" (OUP 1995) as a condition of the emergence of multicellularity under environmental pressure

Scientific achievements
We divide this research axis, which relies on the study of partial differential equations for space and time organisation of biological populations, according to various applications using the same type of mathematical formalisms and methodologies: asymptotic analysis, weak solutions, numerical algorithms.

Aggregation equation
In the mathematical study of collective behavior, an important class of models is given by the aggregation equation. In the presence of a non-smooth interaction potential, solutions of such systems may blow up in finite time. To overcome this difficulty, we have defined weak measure-valued solutions in the sense of duality and its equivalence with gradient flows and entropy solutions in one dimension [117]. The extension to higher dimensions has been studied in [87]. An interesting consequence of this approach is the possibility to use the traditional finite volume approach to design numerical schemes able to capture the good behavior of such weak measure-valued solutions [109], [116].

Identification of the mechanisms of single cell motion
In this research axis, we aim to study the mechanisms of single cell adhesion-based and adhesion free motion. This work is done in the frame of the recently created associated team MaMoCeMa (see Section 9) with the WPI, Vienna. In a first direction [150] with N. Sfakianakis (Heidelberg University), we extended the live-cell motility Filament Based Lamellipodium Model to incorporate the forces exerted on the lamellipodium of the cells due to cell-cell collision and cadherin induced cell-cell adhesion. We took into account the nature of these forces via physical and biological constraints and modelling assumptions. We investigated the effect these new components had in the migration and morphology of the cells through particular experiments. We exhibit moreover the similarities between our simulated cells and HeLa cancer cells.

In a second work done in collaboration with the group of biologist at IST (led by **Michael Sixt** Austria), we developed and analyzed a two-dimensional mathematical model for cells migrating without adhesion capabilities [118]. Cells are represented by their cortex, which is modelled as an elastic curve, subject to an internal pressure force. Net polymerization or depolymerization in the cortex is modelled via local addition or removal of material, driving a cortical flow. The model takes the form of a fully nonlinear degenerate parabolic system. An existence analysis is carried out by adapting ideas from the theory of gradient flows. Numerical simulations show that these simple rules can account for the behavior observed in experiments, suggesting a possible mechanical mechanism for adhesion-independent motility.

Free boundary problems for tumor growth
Fluid dynamic equations are now commonly used to describe tumor growth with two main classes of models: those which describe tumor growth through the dynamics of the density of tumoral cells subjected to a mechanical stress; those describing the tumor through the dynamics of its geometrical domain thanks to a Hele-Shaw-type free boundary model. The first link between these two classes of models has been rigorously obtained thanks to an incompressible limit in [137] for a simple model. This result has motivated the use of another strategy based on viscosity solutions, leading to similar results, in [121].

Since more realistic systems are used in the analysis of medical images, we have extended these studies to include active motion of cells in [136], viscosity in [141] and proved regularity results in [129]. The limiting Hele-Shaw free boundary model has been used to describe mathematically the invasion capacity of a tumour by looking for travelling wave solutions, in [140], see also Methodological axis 3. It is a fundamental but difficult issue to explain rigorously the emergence of instabilities in the direction transversal to the wave propagation. For a simplified model, a complete explanation is obtained in [122].

Two-way coupling of diffusion and growth

We are currently developing a mathematical framework for diffusion equations on time-evolving manifolds, where the evolution of the manifold is a function of the distribution of the diffusing quantity. The need for such a framework takes it roots in developmental biology. Indeed, the growth of an organism is triggered by signaling molecules called morphogens that diffuse in the organism during its development. Meanwhile, the diffusion of the morphogens is itself affected by the changes in shape and size of the organism. In other words, there is a complete coupling between the diffusion of the morphogens and the evolution of the shapes. In addition to the elaboration of this theoretical framework, we also collaborate with a team of developmental biologists from Rutgers University (Camden, New Jersey) to develop a model for the diffusion of Gurken during the oogenesis of *Drosophila*.

Migration of cells in extracellular matrix

A single cell based model has been developed that reproduces a large set of experimental observations of cells migrating in extracellular matrix based on physical mechanisms with mimimal internal cell dynamics. This includes individually migrating cells in micro-channels of different size, and their collective dynamics in case of many cells, as well as the impact of cell division and growth. The model explicitly mimics the extracellular matrix as the cells as deformable objects with explicit filopodia.

Collaborations

- Shanghai Jiao Tong University, joint publications with Min Tang on bacterial models for chemotaxis and free boundary problems for tumor growth.
- Imperial College London, joint works with José Antonio Carrillo on aggregation equation.
- University of Maryland at College Park, UCLA, Univ. of Chicago, Univ. Autónoma de Madrid, Univ. of St. Andrews (Scotland), joint works on mathematics of tumor growth models.
- Joint work with Francesco Rossi (Università di Padova, Italy) and Benedetto Piccoli (Rutgers University, Camden, New Jersey, USA) on Developmental PDEs.
- Cooperation with Shugo Yasuda (University of Hyogo, Kobe, Japan) and Vincent Calvez (EPI Dracula) on the subject of bacterial motion.
- Cooperation with Nathalie Ferrand (INSERM), Michèle Sabbah (INSERM) and Guillaume Vidal (Centre de Recherche Paul Pascal, Bordeaux) on cell aggregation by chemotaxis.
- Nicolas Vauchelet, Université Paris 13

3.4. Methodological axis 3: Model and parameter identification combining stochastic and deterministic approaches in nonlocal and multi-scale models

Personnel

Marie Doumic, Dirk Drasdo.

Project-team positioning

Mamba developed and addressed model and parameter identification methods and strategies in a number of mathematical and computational model applications including growth and fragmentation processes emerging in bacterial growth and protein misfolding, in liver regeneration [103], TRAIL treatment of HeLa cells [74], growth of multicellular spheroids [115], blood detoxification after drug-induced liver damage [149], [107].

This naturally led to increasingly combine methods from various fields: image analysis, statistics, probability, numerical analysis, PDEs, ODEs, agent-based modeling methods, involving inverse methods as well as direct model and model parameter identification in biological and biomedical applications. Model types comprise agent-based simulations for which Mamba is among the leading international groups, and Pharmocokinetic (PK) simulations that have recently combined in integrated models (PhD theses Géraldine Cellière, Noémie Boissier). The challenges related with the methodological variability has led to very fruitful collaborations with internationally renowned specialists of these fields, e.g. for bacterial growth and protein misfolding with Marc Hoffmann (Paris Dauphine) and Patricia Reynaud-Bouret (University of Nice) in statistics, with Tom Banks (Raleigh, USA) and Philippe Moireau (Inria M3DISIM) in inverse problems and data assimilation, and with numerous experimentalists.

Scientific achievements

Direct parameter identification is a great challenge particularly in living systems in which part of parameters at a certain level are under control of processes at smaller scales.

Estimation methods for growing and dividing populations

In this domain, all originated in two papers in collaboration with J.P. Zubelli in 2007 [143], [101], whose central idea was to used the asymptotic steady distribution of the individuals to estimate the division rate. A series of papers improved and extended these first results while keeping the deterministic viewpoint, lastly [78]. The last developments now tackle the still more involved problem of estimating not only the division rate but also the fragmentation kernel (i.e., how the sizes of the offspring are related to the size of the dividing individual) [97]. In parallel, in a long-run collaboration with statisticians, we studied the Piecewise Deterministic Markov Process (PDMP) underlying the equation, and estimated the division rate directly on sample observations of the process, thus making a bridge between the PDE and the PDMP approach in [100], a work which inspired also very recently other groups in statistics and probability [75], [112] and was the basis for Adélaïde Olivier's Ph.D thesis [132], [114] and of some of her more recent works [133][46] (see also axis 5).

Data assimilation and stochastic modeling for protein aggregation

Estimating reaction rates and size distributions of protein polymers is an important step for understanding the mechanisms of protein misfolding and aggregation (see also axis 5). In [63], we settled a framework problem when the experimental measurements consist in the time-dynamics of a moment of the population.

To model the intrinsic variability among experimental curves in aggregation kinetics - an important and poorly understood phenomenon - Sarah Eugène's Ph.D, co-supervised by P. Robert [105], was devoted to the stochastic modeling and analysis of protein aggregation, compared both with the deterministic approach traditionally developed in Mamba [147] and with experiments.

Collaborations

- **Marc Hoffmann,** Université Paris-Dauphine, for the statistical approach to growth and division processes [100], **M. Escobedo**, Bilbao and **M. Tournus**, Marseille, for the deterministic approach.
- **Philippe Moireau,** Inria M3DISIM, for the inverse problem and data assimilation aspects [69], [62]

MATHNEURO Project-Team

3. Research Program

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the long term goal to understand how brain is working. But, beyond the comprehension of brain or even of simpler neural systems in less evolved animals, there is also the desire to exhibit general mechanisms or principles at work in the nervous system. One possible strategy is to propose mathematical models of neural activity, at different space and time scales, depending on the type of phenomena under consideration. However, beyond the mere proposal of new models, which can rapidly result in a plethora, there is also a need to understand some fundamental keys ruling the behaviour of neural networks, and, from this, to extract new ideas that can be tested in real experiments. Therefore, there is a need to make a thorough analysis of these models. An efficient approach, developed in our team, consists of analysing neural networks as dynamical systems. This allows to address several issues. A first, natural issue is to ask about the (generic) dynamics exhibited by the system when control parameters vary. This naturally leads to analyse the bifurcations [70] [71] occurring in the network and which phenomenological parameters control these bifurcations. Another issue concerns the interplay between the neuron dynamics and the synaptic network structure.

3.2. Mean-field and stochastic approaches

Modeling neural activity at scales integrating the effect of thousands of neurons is of central importance for several reasons. First, most imaging techniques are not able to measure individual neuron activity (microscopic scale), but are instead measuring mesoscopic effects resulting from the activity of several hundreds to several hundreds of thousands of neurons. Second, anatomical data recorded in the cortex reveal the existence of structures, such as the cortical columns, with a diameter of about $50\mu m$ to $1mm$, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.
Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate these equations; see Figure 1 . These methods use tools from advanced probability theory such as the theory of Large Deviations [59] and the study of interacting diffusions [3].

3.3. Neural fields

Neural fields are a phenomenological way of describing the activity of population of neurons by delayed integro-differential equations. This continuous approximation turns out to be very useful to model large brain areas such as those involved in visual perception. The mathematical properties of these equations and their solutions are still imperfectly known, in particular in the presence of delays, different time scales and noise.

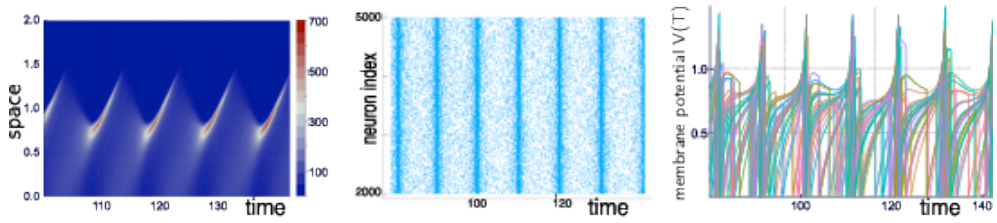


Figure 1. Simulations of the quasi-synchronous state of a stochastic neural network with $N = 5000$ neurons. Left: empirical distribution of membrane potential as a function (t, v). Middle: (raster plot) spiking times as a function of neuron index and time. Right: several membrane potentials $v_i(t)$ as a function of time for $i \in [1, 100]$. Simulated with the Julia Package PDMP.jl from [17]. This figure has been slightly modified from [11].

Our group is developing mathematical and numerical methods for analysing these equations. These methods are based upon techniques from mathematical functional analysis, bifurcation theory [14], [72], equivariant bifurcation analysis, delay equations, and stochastic partial differential equations. We have been able to characterize the solutions of these neural fields equations and their bifurcations, apply and expand the theory to account for such perceptual phenomena as edge, texture [50], and motion perception. We have also developed a theory of the delayed neural fields equations, in particular in the case of constant delays and propagation delays that must be taken into account when attempting to model large size cortical areas [16], [73]. This theory is based on center manifold and normal forms ideas [15].

3.4. Slow-fast dynamics in neuronal models

Neuronal rhythms typically display many different timescales, therefore it is important to incorporate this slow-fast aspect in models. We are interested in this modeling paradigm where slow-fast point models, using Ordinary Differential Equations (ODEs), are investigated in terms of their bifurcation structure and the patterns of oscillatory solutions that they can produce. To insight into the dynamics of such systems, we use a mix of theoretical techniques — such as geometric desingularisation and centre manifold reduction [63] — and numerical methods such as pseudo-arclength continuation [54]. We are interested in families of complex oscillations generated by both mathematical and biophysical models of neurons. In particular, so-called *mixed-mode oscillations (MMOs)* [9], [52], [62], which represent an alternation between subthreshold and spiking behaviour, and *bursting oscillations* [53], [60], also corresponding to experimentally observed behaviour [51]; see Figure 2 . We are working on extending these results to spatio-temporal neural models [2].

3.5. Modeling neuronal excitability

Excitability refers to the all-or-none property of neurons [58], [61]. That is, the ability to respond nonlinearly to an input with a dramatic change of response from "none" — no response except a small perturbation that returns to equilibrium — to "all" — large response with the generation of an action potential or spike before the neuron returns to equilibrium. The return to equilibrium may also be an oscillatory motion of small amplitude; in this case, one speaks of resonator neurons as opposed to integrator neurons. The combination of a spike followed by subthreshold oscillations is then often referred to as mixed-mode oscillations (MMOs) [52]. Slow-fast ODE models of dimension at least three are well capable of reproducing such complex neural oscillations. Part of our research expertise is to analyse the possible transitions between different complex oscillatory patterns of this sort upon input change and, in mathematical terms, this corresponds to understanding the bifurcation structure of the model. Furthermore, the shape of time series of this sort with a given oscillatory pattern can be analysed within the mathematical framework of dynamic bifurcations; see the section on slow-fast dynamics in Neuronal Models. The main example of abnormal neuronal excitability is hyperexcitability

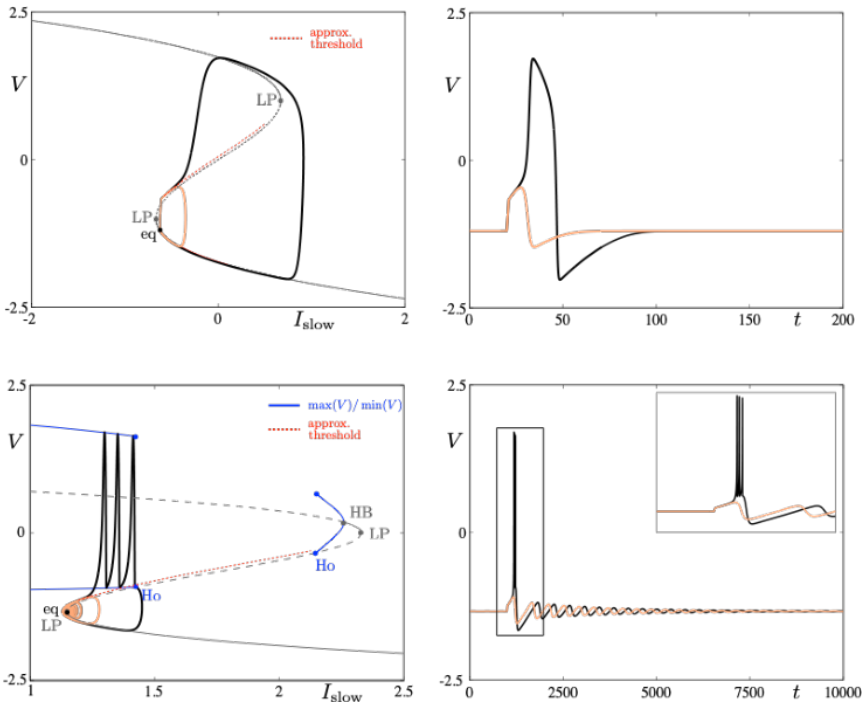


Figure 2. Excitability threshold as slow manifolds in a simple spiking model, namely the FitzHugh-Nagumo model, (top panels) and in a simple bursting model, namely the Hindmarsh-Rose model (bottom panels). This figure is unpublished.

and it is important to understand the biological factors which lead to such excess of excitability and to identify (both in detailed biophysical models and reduced phenomenological ones) the mathematical structures leading to these anomalies. Hyperexcitability is one important trigger for pathological brain states related to various diseases such as chronic migraine [65], epilepsy [75] or even Alzheimer's Disease [64]. A central central axis of research within our group is to revisit models of such pathological scenarios, in relation with a combination of advanced mathematical tools and in partnership with biological labs.

3.6. Synaptic Plasticity

Neural networks show amazing abilities to evolve and adapt, and to store and process information. These capabilities are mainly conditioned by plasticity mechanisms, and especially synaptic plasticity, inducing a mutual coupling between network structure and neuron dynamics. Synaptic plasticity occurs at many levels of organization and time scales in the nervous system [49]. It is of course involved in memory and learning mechanisms, but it also alters excitability of brain areas and regulates behavioral states (e.g., transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge.

Our group is developing mathematical and numerical methods to analyse this mutual interaction. On the one hand, we have shown that plasticity mechanisms [8], [13], Hebbian-like or STDP, have strong effects on neuron dynamics complexity, such as synaptic and propagation delays [16], dynamics complexity reduction, and spike statistics.

MIMESIS Team

3. Research Program

3.1. Real-time computational models for interactive applications

The principal objective of this challenge is to improve, at the numerical level, the efficiency, robustness, and quality of the simulations (see Fig. 2). An important part of our research is dedicated to the development of computational models that remain compatible with real-time computation, i.e., which allow immediate visual or haptic feedback. This typically requires computation times below $50ms$ and in some cases around $1ms$. Such advanced models can not only increase the realism of future training systems, but also act as a bridge toward the development of patient-specific solutions for computer-aided interventions. Additionally, such simulations should run on (high-end) consumer level computers (i.e. with a single multi-core CPU and a dedicated GPU). To reach these goals, we are investigating novel finite element techniques able to cope with complex, potentially ill-defined input data. After developing Smoothed FEM for real-time simulations, we are developing meshless techniques and immersed boundary methods. The first one is well suited for topological changes, which we sometimes need to account for in our simulations. The second is expected to lead to more stable, and numerically efficient, formulations of the finite element method. We are also developing numerical techniques to compute the complex interactions that can take place between anatomical structures or between medical devices and organs. Boundary conditions are known to also play an important role in the solution of such problems. Therefore we are investigating solutions to both identify and model the interactions that take place between the structure of interest and its anatomical environment.

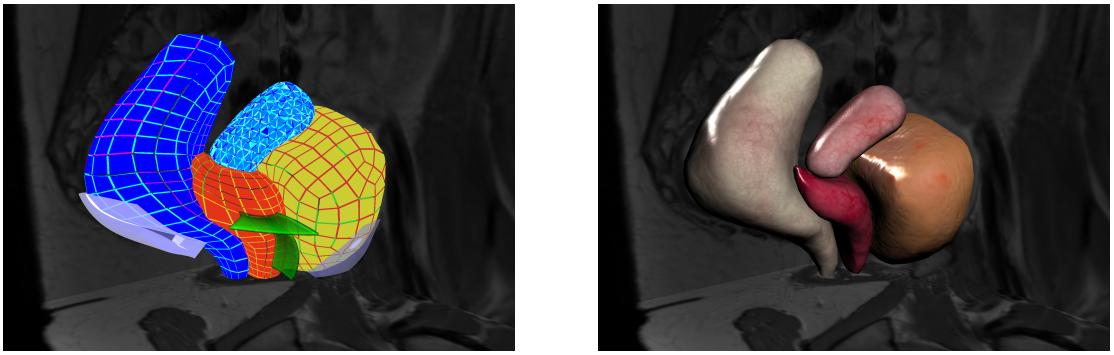


Figure 2. Model of the pelvis with (left) the finite element models of different anatomical structures and (right) their visual representations. Complex interactions take place between these deformable structures. The simulation is computed at interactive rates

3.2. Data-driven simulations

Data-driven simulation has been a recent area of research in our team (see Fig. 3). We have demonstrated that it has the potential to bridge the gap between medical imaging and clinical routine by adapting pre-operative data to the time of the procedure. In the areas of non-rigid registration and augmented reality during surgery, we have demonstrated the benefit of our physics-based approaches with several key publications in major conferences (MICCAI, CVPR, IPCAI, ISMAR).

We have continued this work with an **emphasis on robustness to uncertainty and outliers** in the information extracted in real-time from image data, as well as real-time parameter estimation. This is currently done by **combining Bayesian methods with advanced physics-based methods** to handle uncertainties in image-driven simulations (MICCAI 2017, CVCS 2018).

Finally, Bayesian or similar methods require to perform a large amount of simulations to sample the domain space, even when using efficient methods such as Reduced Order Unscented Kalman Filters. For this reason, we are investigating the use of neural networks to perform predictions instead of using full numerical simulations. Our latest paper [22] at MICCAI 2019 shows it is possible to **teach a neural network from numerical simulations** and **predict**, with good accuracy, **the deformation of an organ**.

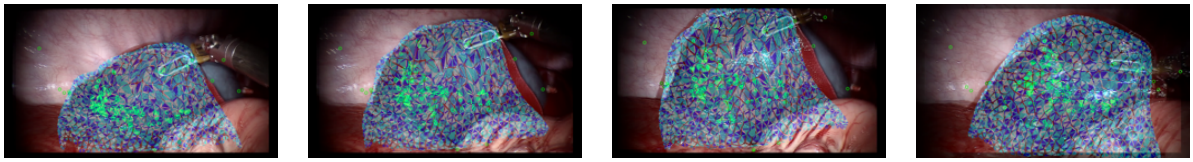


Figure 3. Real-time deformation of a virtual liver according to tissue motion tracked in laparoscopic images.

MNEMOSYNE Project-Team

3. Research Program

3.1. Integrative and Cognitive Neuroscience

The human brain is often considered as the most complex system dedicated to information processing. This multi-scale complexity, described from the metabolic to the network level, is particularly studied in integrative neuroscience, the goal of which is to explain how cognitive functions (ranging from sensorimotor coordination to executive functions) emerge from (are the result of the interaction of) distributed and adaptive computations of processing units, displayed along neural structures and information flows. Indeed, beyond the astounding complexity reported in physiological studies, integrative neuroscience aims at extracting, in simplifying models, regularities at various levels of description. From a mesoscopic point of view, most neuronal structures (and particularly some of primary importance like the cortex, cerebellum, striatum, hippocampus) can be described through a regular organization of information flows and homogenous learning rules, whatever the nature of the processed information. From a macroscopic point of view, the arrangement in space of neuronal structures within the cerebral architecture also obeys a functional logic, the sketch of which is captured in models describing the main information flows in the brain, the corresponding loops built in interaction with the external and internal (bodily and hormonal) world and the developmental steps leading to the acquisition of elementary sensorimotor skills up to the most complex executive functions.

In summary, integrative neuroscience builds, on an overwhelming quantity of data, a simplifying and interpretative grid suggesting homogenous local computations and a structured and logical plan for the development of cognitive functions. They arise from interactions and information exchange between neuronal structures and the external and internal world and also within the network of structures.

This domain is today very active and stimulating because it proposes, of course at the price of simplifications, global views of cerebral functioning and more local hypotheses on the role of subsets of neuronal structures in cognition. In the global approaches, the integration of data from experimental psychology and clinical studies leads to an overview of the brain as a set of interacting memories, each devoted to a specific kind of information processing [53]. It results also in longstanding and very ambitious studies for the design of cognitive architectures aiming at embracing the whole cognition. With the notable exception of works initiated by [50], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [55], [44]. During the recent years this domain has gained an increasing interest in neuroscience and has become an essential tool for scientific developments in most fields in neuroscience, from the molecule to the system. In this view, all the objectives of our team can be described as possible progresses in computational neuroscience. Accordingly, it can be underlined that the systemic view that we promote can offer original contributions in the sense that, whereas most classical models in computational neuroscience focus on the better understanding of the structure/function relationship for isolated specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogenous modes of computing, which is rarely addressed in classical computational neuroscience.

We also insist on another aspect of computational neuroscience which is, in our opinion, at the core of the involvement of computer scientists and mathematicians in the domain and on which we think we could particularly contribute. Indeed, we think that our primary abilities in numerical sciences imply that our developments are characterized above all by the effectiveness of the corresponding computations: We provide biologically inspired architectures with effective computational properties, such as robustness to noise, self-organization, on-line learning. We more generally underline the requirement that our models must also mimick biology through its most general law of homeostasis and self-adaptability in an unknown and changing environment. This means that we propose to numerically experiment such models and thus provide effective methods to falsify them.

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [41], [49] and were the basis for several contributions in our group [54], [51], [56]. Ensuring such a rigor in the computations associated to our systemic and large scale approach is of central importance.

Equally important is the choice for the formalism of computation, extensively discussed in the connectionist domain. Spiking neurons are today widely recognized of central interest to study synchronization mechanisms and neuronal coupling at the microscopic level [42]; the associated formalism [47] can be possibly considered for local studies or for relating our results with this important domain in connectionism. Nevertheless, we remain mainly at the mesoscopic level of modeling, the level of the neuronal population, and consequently interested in the formalism developed for dynamic neural fields [39], that demonstrated a richness of behavior [43] adapted to the kind of phenomena we wish to manipulate at this level of description. Our group has a long experience in the study and adaptation of the properties of neural fields [51], [52] and their use for observing the emergence of typical cortical properties [46]. In the envisioned development of more complex architectures and interplay between structures, the exploration of mathematical properties such as stability and boundedness and the observation of emerging phenomena is one important objective. This objective is also associated with that of capitalizing our experience and promoting good practices in our software production. In summary, we think that this systemic approach also brings to computational neuroscience new case studies where heterogenous and adaptive models with various time scales and parameters have to be considered jointly to obtain a mastered substratum of computation. This is particularly critical for large scale deployments.

3.3. Machine Learning

The adaptive properties of the nervous system are certainly among its most fascinating characteristics, with a high impact on our cognitive functions. Accordingly, machine learning is a domain [48] that aims at giving such characteristics to artificial systems, using a mathematical framework (probabilities, statistics, data analysis, etc.). Some of its most famous algorithms are directly inspired from neuroscience, at different levels. Connectionist learning algorithms implement, in various neuronal architectures, weight update rules, generally derived from the hebbian rule, performing non supervised (e.g. Kohonen self-organizing maps), supervised (e.g. layered perceptrons) or associative (e.g. Hopfield recurrent network) learning. Other algorithms, not necessarily connectionist, perform other kinds of learning, like reinforcement learning. Machine learning is a very mature domain today and all these algorithms have been extensively studied, at both the theoretical and practical levels, with much success. They have also been related to many functions (in the living and artificial domains) like discrimination, categorisation, sensorimotor coordination, planning, etc. and several neuronal structures have been proposed as the substratum for these kinds of learning [45], [38]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

At the cognitive level, most of the problems we face do not rely on only one kind of learning and require instead skills that have to be learned in preliminary steps. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead of the classical view in machine learning of a flat architecture, a more complex network of modules must be

considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episods and supervising the training of a procedural memory in the cortex.

At the behavioral level, as mentionned above, our systemic approach underlines the fundamental links between the adaptive system and the internal and external world. The internal world includes proprioception and interoception, giving information about the body and its needs for integrity and other fundamental programs. The external world includes physical laws that have to be learned and possibly intelligent agents for more complex interactions. Both involve sensors and actuators that are the interfaces with these worlds and close the loops. Within this rich picture, machine learning generally selects one situation that defines useful sensors and actuators and a corpus with properly segmented data and time, and builds a specific architecture and its corresponding criteria to be satisfied. In our approach however, the first question to be raised is to discover what is the goal, where attention must be focused on and which previous skills must be exploited, with the help of a dynamic architecture and possibly other partners. In this domain, the behavioral and the developmental sciences, observing how and along which stages an agent learns, are of great help to bring some structure to this high dimensional problem.

At the implementation level, this analysis opens many fundamental challenges, hardly considered in machine learning : stability must be preserved despite on-line continuous learning; criteria to be satisfied often refer to behavioral and global measurements but they must be translated to control the local circuit level; in an incremental or developmental approach, how will the development of new functions preserve the integrity and stability of others? In addition, this continous re-arrangement is supposed to involve several kinds of learning, at different time scales (from msec to years in humans) and to interfer with other phenomena like variability and meta-plasticity.

In summary, our main objective in machine learning is to propose on-line learning systems, where several modes of learning have to collaborate and where the protocoles of training are realistic. We promote here a *really autonomous* learning, where the agent must select by itself internal resources (and build them if not available) to evolve at the best in an unknown world, without the help of any *deus-ex-machina* to define parameters, build corpus and define training sessions, as it is generally the case in machine learning. To that end, autonomous robotics (*cf.* § 3.4) is a perfect testbed.

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges from interactions of the body in direct loops with the world (*embodiment of cognition* [40]). In addition to real robotic platforms, software implementations of autonomous robotic systems including components dedicated to their body and their environment will be also possibly exploited, considering that they are also a tool for studying conditions for a real autonomous learning.

A real autonomy can be obtained only if the robot is able to define its goal by itself, without the specification of any high level and abstract cost function or rewarding state. To ensure such a capability, we propose to endow the robot with an artificial physiology, corresponding to perceive some kind of pain and pleasure. It may consequently discriminate internal and external goals (or situations to be avoided). This will mimick circuits related to fundamental needs (e.g. hunger and thirst) and to the preservation of bodily integrity. An important objective is to show that more abstract planning capabilities can arise from these basic goals.

A real autonomy with an on-line continuous learning as described in § 3.3 will be made possible by the elaboration of protocols of learning, as it is the case, in animal conditioning, for experimental studies where performance on a task can be obtained only after a shaping in increasingly complex tasks. Similarly,

developmental sciences can teach us about the ordered elaboration of skills and their association in more complex schemes. An important challenge here is to translate these hints at the level of the cerebral architecture.

As a whole, autonomous robotics permits to assess the consistency of our models in realistic condition of use and offers to our colleagues in behavioral sciences an object of study and comparison, regarding behavioral dynamics emerging from interactions with the environment, also observable at the neuronal level.

In summary, our main contribution in autonomous robotics is to make autonomy possible, by various means corresponding to endow robots with an artificial physiology, to give instructions in a natural and incremental way and to prioritize the synergy between reactive and robust schemes over complex planning structures.

MONC Project-Team

3. Research Program

3.1. Introduction

We are working in the context of data-driven medicine against cancer. We aim at coupling mathematical models with data to address relevant challenges for biologists and clinicians in order for instance to improve our understanding in cancer biology and pharmacology, assist the development of novel therapeutic approaches or develop personalized decision-helping tools for monitoring the disease and evaluating therapies.

More precisely, our research on mathematical oncology is three-fold:

- Axis 1: Tumor modeling for patient-specific simulations: *Clinical monitoring. Numerical markers from imaging data. Radiomics*.
- Axis 2: Bio-physical modeling for personalized therapies: *Electroporation from cells to tissue. Radiotherapy*.
- Axis 3: Quantitative cancer modeling for biological, clinical and preclinical studies: *Biological mechanisms. Metastatic dissemination. Pharmacometrics*.

In the first axis, we aim at producing patient-specific simulations of the growth of a tumor or its response to treatment starting from a series of images. We hope to be able to offer a valuable insight on the disease to the clinicians in order to improve the decision process. This would be particularly useful in the cases of relapses or for metastatic diseases.

The second axis aims at modeling biophysical therapies like electroporation, but also radiotherapy, thermo-ablations, radio-frequency ablations or electroporation that play a crucial role for a local treatment of the disease if possible limiting the metastatic dissemination, which is precisely the clinical context where the techniques of axis 1 will be applied.

The third axis is essential since it is a way to better understand and model the biological reality of cancer growth and the (possibly complex) effects of therapeutic intervention. Modeling in this case also helps to interpret the experimental results and improve the accuracy of the models used in Axis 1. Technically speaking, some of the computing tools are similar to those of Axis 1.

3.2. Axis 1: Tumor modeling for patient-specific simulations

The gold standard treatment for most cancers is surgery. In the case where total resection of the tumor is possible, the patient often benefits from an adjuvant therapy (radiotherapy, chemotherapy, targeted therapy or a combination of them) in order to eliminate the potentially remaining cells that may not be visible. In this case personalized modeling of tumor growth is useless and statistical modeling will be able to quantify the risk of relapse, the mean progression-free survival time...However if total resection is not possible or if metastases emerge from distant sites, clinicians will try to control the disease for as long as possible. A wide set of tools are available. Clinicians may treat the disease by physical interventions (radiofrequency ablation, cryoablation, radiotherapy, electroporation, focalized ultrasound,...) or chemical agents (chemotherapies, targeted therapies, antiangiogenic drugs, immunotherapies, hormonotherapies). One can also decide to monitor the patient without any treatment (this is the case for slowly growing tumors like some metastases to the lung, some lymphomas or for some low grade glioma). A reliable patient-specific model of tumor evolution with or without therapy may have different uses:

- Case without treatment: the evaluation of the growth of the tumor would offer a useful indication for the time at which the tumor may reach a critical size. For example, radiofrequency ablation of pulmonary lesion is very efficient as long as the diameter of the lesion is smaller than 3 cm. Thus, the prediction can help the clinician plan the intervention. For slowly growing tumors, quantitative

modeling can also help to decide at what time interval the patient has to undergo a CT-scan. CT-scans are irradiative exams and there is a challenge for decreasing their occurrence for each patient. It has also an economical impact. And if the disease evolution starts to differ from the prediction, this might mean that some events have occurred at the biological level. For instance, it could be the rise of an aggressive phenotype or cells that leave a dormancy state. This kind of events cannot be predicted, but some mismatch with respect to the prediction can be an indirect proof of their existence. It could be an indication for the clinician to start a treatment.

- Case with treatment: a model can help to understand and to quantify the final outcome of a treatment using the early response. It can help for a redefinition of the treatment planning. Modeling can also help to anticipate the relapse by analyzing some functional aspects of the tumor. Again, a deviation with respect to reference curves can mean a lack of efficiency of the therapy or a relapse. Moreover, for a long time, the response to a treatment has been quantified by the RECIST criteria which consists in (roughly speaking) measuring the diameters of the largest tumor of the patient, as it is seen on a CT-scan. This criteria is still widely used and was quite efficient for chemotherapies and radiotherapies that induce a decrease of the size of the lesion. However, with the systematic use of targeted therapies and anti-angiogenic drugs that modify the physiology of the tumor, the size may remain unchanged even if the drug is efficient and deeply modifies the tumor behavior. One better way to estimate this effect could be to use functional imaging (Pet-scan, perfusion or diffusion MRI, ...), a model can then be used to exploit the data and to understand in what extent the therapy is efficient.
- Optimization: currently, we do not believe that we can optimize a particular treatment in terms of distribution of doses, number, planning with the model that we will develop in a medium term perspective.

The scientific challenge is therefore as follows: given the history of the patient, the nature of the primitive tumor, its histopathology, knowing the treatments that patients have undergone, some biological facts on the tumor and having a sequence of images (CT-scan, MRI, PET or a mix of them), are we able to provide a numerical simulation of the extension of the tumor and of its metabolism that fits as best as possible with the data (CT-scans or functional data) and that is predictive in order to address the clinical cases described above?

Our approach relies on the elaboration of PDE models and their parametrization with images by coupling deterministic and stochastic methods. The PDE models rely on the description of the dynamics of cell populations. The number of populations depends on the pathology. For example, for glioblastoma, one needs to use proliferative cells, invasive cells, quiescent cells as well as necrotic tissues to be able to reproduce realistic behaviors of the disease. In order to describe the relapse for hepatic metastases of gastro-intestinal stromal tumor (gist), one needs three cell populations: proliferative cells, healthy tissue and necrotic tissue.

The law of proliferation is often coupled with a model for the angiogenesis. However such models of angiogenesis involve too many non measurable parameters to be used with real clinical data and therefore one has to use simplified or even simplistic versions. The law of proliferation often mimics the existence of an hypoxia threshold, it consists of an ODE. or a PDE that describes the evolution of the growth rate as a combination of sigmoid functions of nutrients or roughly speaking oxygen concentration. Usually, several laws are available for a given pathology since at this level, there are no quantitative argument to choose a particular one.

The velocity of the tumor growth differs depending on the nature of the tumor. For metastases, we will derive the velocity thanks to Darcy's law in order to express that the extension of the tumor is basically due to the increase of volume. This gives a sharp interface between the metastasis and the surrounding healthy tissues, as observed by anatomopathologists. For primitive tumors like glioma or lung cancer, we use reaction-diffusion equations in order to describe the invasive aspects of such primitive tumors.

The modeling of the drugs depends on the nature of the drug: for chemotherapies, a death term can be added into the equations of the population of cells, while antiangiogenic drugs have to be introduced in a angiogenic model. Resistance to treatment can be described either by several populations of cells or with non-constant growth or death rates. As said before, it is still currently difficult to model the changes of phenotype

or mutations, we therefore propose to investigate this kind of phenomena by looking at deviations of the numerical simulations compared to the medical observations.

The calibration of the model is achieved by using a series (at least 2) of images of the same patient and by minimizing a cost function. The cost function contains at least the difference between the volume of the tumor that is measured on the images with the computed one. It also contains elements on the geometry, on the necrosis and any information that can be obtained through the medical images. We will pay special attention to functional imaging (PET, perfusion and diffusion MRI). The inverse problem is solved using a gradient method coupled with some Monte-Carlo type algorithm. If a large number of similar cases is available, one can imagine to use statistical algorithms like random forests to use some non quantitative data like the gender, the age, the origin of the primitive tumor...for example for choosing the model for the growth rate for a patient using this population knowledge (and then to fully adapt the model to the patient by calibrating this particular model on patient data) or for having a better initial estimation of the modeling parameters. We have obtained several preliminary results concerning lung metastases including treatments and for metastases to the liver.

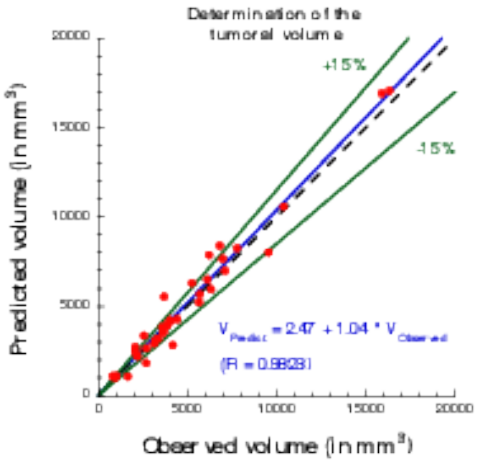


Figure 4. Plot showing the accuracy of our prediction on meningioma volume. Each point corresponds to a patient whose two first exams were used to calibrate our model. A patient-specific prediction was made with this calibrated model and compared with the actual volume as measured on a third time by clinicians. A perfect prediction would be on the black dashed line. Medical data was obtained from Prof. Loiseau, CHU Pellegrin.

3.3. Axis 2: Bio-physical modeling for personalized therapies

In this axis, we investigate locoregional therapies such as radiotherapy, irreversible electroporation. Electroporation consists in increasing the membrane permeability of cells by the delivery of high voltage pulses. This non-thermal phenomenon can be transient (reversible) or irreversible (IRE). IRE or electro-chemotherapy – which is a combination of reversible electroporation with a cytotoxic drug – are essential tools for the treatment of a metastatic disease. Numerical modeling of these therapies is a clear scientific challenge. Clinical applications of the modeling are the main target, which thus drives the scientific approach, even though theoretical studies in order to improve the knowledge of the biological phenomena, in particular for electroporation, should also be addressed. However, this subject is quite wide and we focus on two particular approaches: some aspects of radiotherapies and electro-chemotherapy. This choice is motivated partly by pragmatic reasons: we

already have collaborations with physicians on these therapies. Other treatments could be probably treated with the same approach, but we do not plan to work on this subject on a medium term.

- Radiotherapy (RT) is a common therapy for cancer. Typically, using a CT scan of the patient with the structures of interest (tumor, organs at risk) delineated, the clinicians optimize the dose delivery to treat the tumor while preserving healthy tissues. The RT is then delivered every day using low resolution scans (CBCT) to position the beams. Under treatment the patient may lose weight and the tumor shrinks. These changes may affect the propagation of the beams and subsequently change the dose that is effectively delivered. It could be harmful for the patient especially if sensitive organs are concerned. In such cases, a replanification of the RT could be done to adjust the therapeutical protocol. Unfortunately, this process takes too much time to be performed routinely. The challenges faced by clinicians are numerous, we focus on two of them:
  - *Detecting the need of replanification:* we are using the positioning scans to evaluate the movement and deformation of the various structures of interest. Thus we can detect whether or not a structure has moved out of the safe margins (fixed by clinicians) and thus if a replanification may be necessary. In a retrospective study, our work can also be used to determine RT margins when there are no standard ones. A collaboration with the RT department of Institut Bergonié is underway on the treatment of retroperitoneal sarcoma and ENT tumors (head and neck cancers). A retrospective study was performed on 11 patients with retro-peritoneal sarcoma. The results have shown that the safety margins (on the RT) that clinicians are currently using are probably not large enough. The tool used in this study was developed by an engineer funded by Inria (Cynthia Périer, ADT Sesar). We used well validated methods from a level-set approach and segmentation / registration methods. The originality and difficulty lie in the fact that we are dealing with real data in a clinical setup. Clinicians have currently no way to perform complex measurements with their clinical tools. This prevents them from investigating the replanification. Our work and the tools developed pave the way for easier studies on evaluation of RT plans in collaboration with Institut Bergonié. *There was no modeling involved in this work that arose during discussions with our collaborators.* The main purpose of the team is to have meaningful outcomes of our research for clinicians, sometimes it implies leaving a bit our area of expertise.
  - *Evaluating RT efficacy and finding correlation between the radiological responses and the clinical outcome:* our goal is to help doctors to identify correlation between the response to RT (as seen on images) and the longer term clinical outcome of the patient. Typically, we aim at helping them to decide when to plan the next exam after the RT. For patients whose response has been linked to worse prognosis, this exam would have to be planned earlier. This is the subject of collaborations with Institut Bergonié and CHU Bordeaux on different cancers (head and neck, pancreas). The response is evaluated from image markers (*e.g.* using texture information) or with a mathematical model developed in Axis 1. The other challenges are either out of reach or not in the domain of expertise of the team. Yet our works may tackle some important issues for adaptive radiotherapy.
- Both IRE and electrochemotherapy are anticancerous treatments based on the same phenomenon: the electroporation of cell membranes. This phenomenon is known for a few decades but it is still not well understood, therefore our interest is two fold:
  1. We want to use mathematical models in order to better understand the biological behavior and the effect of the treatment. We work in tight collaboration with biologists and bioeletromagneticians to derive precise models of cell and tissue electroporation, in the continuity of the research program of the Inria team-project MC2. These studies lead to complex non-linear mathematical models involving some parameters (as less as possible). Numerical methods to compute precisely such models and the calibration of the parameters with the experimental data are then addressed. Tight collaborations with the Vectorology and Anticancerous Therapies (VAT) of IGR at Villejuif, Laboratoire Ampère of Ecole

Centrale Lyon and the Karlsruhe Institute of technology will continue, and we aim at developing new collaborations with Institute of Pharmacology and Structural Biology (IPBS) of Toulouse and the Laboratory of Molecular Pathology and Experimental Oncology (LMPEO) at CNR Rome, in order to understand differences of the electroporation of healthy cells and cancer cells in spheroids and tissues.

2. This basic research aims at providing new understanding of electroporation, however it is necessary to address, particular questions raised by radio-oncologists that apply such treatments. One crucial question is "What pulse or what train of pulses should I apply to electroporate the tumor if the electrodes are located as given by the medical images"? Even if the real-time optimization of the placement of the electrodes for deep tumors may seem quite utopian since the clinicians face too many medical constraints that cannot be taken into account (like the position of some organs, arteries, nerves...), one can expect to produce real-time information of the validity of the placement done by the clinician. Indeed, once the placement is performed by the radiologists, medical images are usually used to visualize the localization of the electrodes. Using these medical data, a crucial goal is to provide a tool in order to compute in real-time and visualize the electric field and the electroporated region directly on theses medical images, to give the doctors a precise knowledge of the region affected by the electric field. In the long run, this research will benefit from the knowledge of the theoretical electroporation modeling, but it seems important to use the current knowledge of tissue electroporation – even quite rough –, in order to rapidly address the specific difficulty of such a goal (real-time computing of non-linear model, image segmentation and visualization). Tight collaborations with CHU Pellegrin at Bordeaux, and CHU J. Verdier at Bondy are crucial.

- Radiofrequency ablation. In a collaboration with Hopital Haut Leveque, CHU Bordeaux we are trying to determine the efficacy and risk of relapse of hepatocellular carcinoma treated by radiofrequency ablation. For this matter we are using geometrical measurements on images (margins of the RFA, distance to the boundary of the organ) as well as texture information to statistically evaluate the clinical outcome of patients.
- Intensity focused ultrasound. In collaboration with Utrecht Medical center, we aim at tackling several challenges in clinical applications of IFU: target tracking, dose delivery...

3.4. Axis 3: Quantitative cancer modeling for biological and preclinical studies

With the emergence and improvement of a plethora of experimental techniques, the molecular, cellular and tissue biology has operated a shift toward a more quantitative science, in particular in the domain of cancer biology. These quantitative assays generate a large amount of data that call for theoretical formalism in order to better understand and predict the complex phenomena involved. Indeed, due to the huge complexity underlying the development of a cancer disease that involves multiple scales (from the genetic, intra-cellular scale to the scale of the whole organism), and a large number of interacting physiological processes (see the so-called "hallmarks of cancer"), several questions are not fully understood. Among these, we want to focus on the most clinically relevant ones, such as the general laws governing tumor growth and the development of metastases (secondary tumors, responsible of 90% of the deaths from a solid cancer). In this context, it is thus challenging to exploit the diversity of the data available in experimental settings (such as *in vitro* tumor spheroids or *in vivo* mice experiments) in order to improve our understanding of the disease and its dynamics, which in turn lead to validation, refinement and better tuning of the macroscopic models used in the axes 1 and 2 for clinical applications.

In recent years, several new findings challenged the classical vision of the metastatic development biology, in particular by the discovery of organism-scale phenomena that are amenable to a dynamical description in terms of mathematical models based on differential equations. These include the angiogenesis-mediated distant inhibition of secondary tumors by a primary tumor the pre-metastatic niche or the self-seeding phenomenon Building a general, cancer type specific, comprehensive theory that would integrate these dynamical processes

remains an open challenge. On the therapeutic side, recent studies demonstrated that some drugs (such as the Sunitinib), while having a positive effect on the primary tumor (reduction of the growth), could *accelerate* the growth of the metastases. Moreover, this effect was found to be scheduling-dependent. Designing better ways to use this drug in order to control these phenomena is another challenge. In the context of combination therapies, the question of the *sequence* of administration between the two drugs is also particularly relevant.

One of the technical challenge that we need to overcome when dealing with biological data is the presence of potentially very large inter-animal (or inter-individual) variability.

Starting from the available multi-modal data and relevant biological or therapeutic questions, our purpose is to develop adapted mathematical models (*i.e.* identifiable from the data) that recapitulate the existing knowledge and reduce it to its more fundamental components, with two main purposes:

1. to generate quantitative and empirically testable predictions that allow to assess biological hypotheses or
2. to investigate the therapeutic management of the disease and assist preclinical studies of anti-cancerous drug development.

We believe that the feedback loop between theoretical modeling and experimental studies can help to generate new knowledge and improve our predictive abilities for clinical diagnosis, prognosis, and therapeutic decision. Let us note that the first point is in direct link with the axes 1 and 2 of the team since it allows us to experimentally validate the models at the biological scale (*in vitro* and *in vivo* experiments) for further clinical applications.

More precisely, we first base ourselves on a thorough exploration of the biological literature of the biological phenomena we want to model: growth of tumor spheroids, *in vivo* tumor growth in mice, initiation and development of the metastases, effect of anti-cancerous drugs. Then we investigate, using basic statistical tools, the data we dispose, which can range from: spatial distribution of heterogeneous cell population within tumor spheroids, expression of cell markers (such as green fluorescent protein for cancer cells or specific antibodies for other cell types), bioluminescence, direct volume measurement or even intra-vital images obtained with specific imaging devices. According to the data type, we further build dedicated mathematical models that are based either on PDEs (when spatial data is available, or when time evolution of a structured density can be inferred from the data, for instance for a population of tumors) or ODEs (for scalar longitudinal data). These models are confronted to the data by two principal means:

1. when possible, experimental assays can give a direct measurement of some parameters (such as the proliferation rate or the migration speed) or
2. statistical tools to infer the parameters from observables of the model.

This last point is of particular relevance to tackle the problem of the large inter-animal variability and we use adapted statistical tools such as the mixed-effects modeling framework.

Once the models are shown able to describe the data and are properly calibrated, we use them to test or simulate biological hypotheses. Based on our simulations, we then aim at proposing to our biological collaborators new experiments to confirm or infirm newly generated hypotheses, or to test different administration protocols of the drugs. For instance, in a collaboration with the team of the professor Andreas Bikfalvi (Laboratoire de l'Angiogénèse et du Micro-environnement des Cancers, Inserm, Bordeaux), based on confrontation of a mathematical model to multi-modal biological data (total number of cells in the primary and distant sites and MRI), we could demonstrate that the classical view of metastatic dissemination and development (one metastasis is born from one cell) was probably inaccurate, in mice grafted with metastatic kidney tumors. We then proposed that metastatic germs could merge or attract circulating cells. Experiments involving cells tagged with two different colors are currently performed in order to confirm or infirm this hypothesis.

Eventually, we use the large amount of temporal data generated in preclinical experiments for the effect of anti-cancerous drugs in order to design and validate mathematical formalisms translating the biological mechanisms of action of these drugs for application to clinical cases, in direct connection with the axis 1. We have a special focus on targeted therapies (designed to specifically attack the cancer cells while sparing the

healthy tissue) such as the Sunitinib. This drug is indeed indicated as a first line treatment for metastatic renal cancer and we plan to conduct a translational study coupled between A. Bikfalvi's laboratory and medical doctors, F. Cornelis (radiologist) and A. Ravaud (head of the medical oncology department).

MORPHEME Project-Team

3. Research Program

3.1. Research program

The recent advent of an increasing number of new microscopy techniques giving access to high throughput screenings and micro or nano-metric resolutions provides a means for quantitative imaging of biological structures and phenomena. To conduct quantitative biological studies based on these new data, it is necessary to develop non-standard specific tools. This requires using a multi-disciplinary approach. We need biologists to define experiment protocols and interpret the results, but also physicists to model the sensors, computer scientists to develop algorithms and mathematicians to model the resulting information. These different expertises are combined within the Morpheme team. This generates a fecund frame for exchanging expertise, knowledge, leading to an optimal framework for the different tasks (imaging, image analysis, classification, modeling). We thus aim at providing adapted and robust tools required to describe, explain and model fundamental phenomena underlying the morphogenesis of cellular and supra-cellular biological structures. Combining experimental manipulations, in vivo imaging, image processing and computational modeling, we plan to provide methods for the quantitative analysis of the morphological changes that occur during development. This is of key importance as the morphology and topology of mesoscopic structures govern organ and cell function. Alterations in the genetic programs underlying cellular morphogenesis have been linked to a range of pathologies.
Biological questions we will focus on include:

1. what are the parameters and the factors controlling the establishment of ramified structures? (Are they really organize to ensure maximal coverage? How are genetic and physical constraints limiting their morphology?),
2. how are newly generated cells incorporated into reorganizing tissues during development? (is the relative position of cells governed by the lineage they belong to?)

Our goal is to characterize different populations or development conditions based on the shape of cellular and supra-cellular structures, e.g. micro-vascular networks, dendrite/axon networks, tissues from 2D, 2D+t, 3D or 3D+t images (obtained with confocal microscopy, video-microscopy, photon-microscopy or micro-tomography). We plan to extract shapes or quantitative parameters to characterize the morphometric properties of different samples. On the one hand, we will propose numerical and biological models explaining the temporal evolution of the sample, and on the other hand, we will statistically analyze shapes and complex structures to identify relevant markers for classification purposes. This should contribute to a better understanding of the development of normal tissues but also to a characterization at the supra-cellular scale of different pathologies such as Alzheimer, cancer, diabetes, or the Fragile X Syndrome. In this multidisciplinary context, several challenges have to be faced. The expertise of biologists concerning sample generation, as well as optimization of experimental protocols and imaging conditions, is of course crucial. However, the imaging protocols optimized for a qualitative analysis may be sub-optimal for quantitative biology. Second, sample imaging is only a first step, as we need to extract quantitative information. Achieving quantitative imaging remains an open issue in biology, and requires close interactions between biologists, computer scientists and applied mathematicians. On the one hand, experimental and imaging protocols should integrate constraints from the downstream computer-assisted analysis, yielding to a trade-off between qualitative optimized and quantitative optimized protocols. On the other hand, computer analysis should integrate constraints specific to the biological problem, from acquisition to quantitative information extraction. There is therefore a need of specificity for embedding precise biological information for a given task. Besides, a level of generality is also desirable for addressing data from different teams acquired with different protocols and/or sensors. The mathematical modeling of the physics of the acquisition system will yield higher performance reconstruction/restoration algorithms in terms of accuracy. Therefore, physicists and computer scientists have to work together. Quantitative information extraction also has to deal with both the complexity of the structures of interest (e.g., very

dense network, small structure detection in a volume, multiscale behavior, ...) and the unavoidable defects of in vivo imaging (artifacts, missing data, ...). Incorporating biological expertise in model-based segmentation methods provides the required specificity while robustness gained from a methodological analysis increases the generality. Finally, beyond image processing, we aim at quantifying and then statistically analyzing shapes and complex structures (e.g., neuronal or vascular networks), static or in evolution, taking into account variability. In this context, learning methods will be developed for determining (dis)similarity measures between two samples or for determining directly a classification rule using discriminative models, generative models, or hybrid models. Besides, some metrics for comparing, classifying and characterizing objects under study are necessary. We will construct such metrics for biological structures such as neuronal or vascular networks. Attention will be paid to computational cost and scalability of the developed algorithms: biological experimentations generally yield huge data sets resulting from high throughput screenings. The research of Morpheme will be developed along the following axes:

- **Imaging:** this includes i) definition of the studied populations (experimental conditions) and preparation of samples, ii) definition of relevant quantitative characteristics and optimized acquisition protocol (staining, imaging, ...) for the specific biological question, and iii) reconstruction/restoration of native data to improve the image readability and interpretation.
- **Feature extraction:** this consists in detecting and delineating the biological structures of interest from images. Embedding biological properties in the algorithms and models is a key issue. Two main challenges are the variability, both in shape and scale, of biological structures and the huge size of data sets. Following features along time will allow to address morphogenesis and structure development.
- **Classification/Interpretation:** considering a database of images containing different populations, we can infer the parameters associated with a given model on each dataset from which the biological structure under study has been extracted. We plan to define classification schemes for characterizing the different populations based either on the model parameters, or on some specific metric between the extracted structures.
- **Modeling:** two aspects will be considered. This first one consists in modeling biological phenomena such as axon growing or network topology in different contexts. One main advantage of our team is the possibility to use the image information for calibrating and/or validating the biological models. Calibration induces parameter inference as a main challenge. The second aspect consists in using a prior based on biological properties for extracting relevant information from images. Here again, combining biology and computer science expertise is a key point.

MOSAIC Project-Team

3. Research Program

3.1. Axis1: Representation of biological organisms and their forms in silico

The modeling of organism development requires a formalization of the concept of form, *i.e.* a mathematical definition of what is a form and how it can change in time, together with the development of efficient algorithms to construct corresponding computational representations from observations, to manipulate them and associate local molecular and physical information with them. Our aim is threefold. First, we will develop new computational structures that make it possible to represent complex forms efficiently in space and time. For branching forms, the challenge will be to reduce the computational burden of the current tree-like representations that usually stems from their exponential increase in size during growth. For tissue structures, we will seek to develop models that integrate seamlessly continuous representations of the cell geometry and discrete representations of their adjacency network in dynamical and adaptive framework. Second, we will explore the use of machine learning strategies to set up robust and adaptive strategies to construct form representations in computers from imaging protocols. Finally, we will develop the notion of digital atlases of development, by mapping patterns of molecular (gene activity, hormones concentrations, cell polarity, ...) and physical (stress, mechanical properties, turgidity, ...) expressions observed at different stages of development on models representing average form development and by providing tools to manipulate and explore these digital atlases.

3.2. Axis2: Data-driven models of form development

Our aim in this second research axis will be to develop models of physiological patterning and bio-physical growth to simulate the development of 3D biological forms in a realistic way. Models of key processes participating to different aspects of morphogenesis (signaling, transport, molecular regulation, cell division, etc.) will be developed and tested *in silico* on 3D data structures reconstructed from digitized forms. The way these component-based models scale-up at more abstract levels where forms can be considered as continuums will also be investigated. Altogether, this will lead us to design first highly integrated models of form development, combining models of different processes in one computational structure representing the form, and to analyze how these processes interact in the course of development to build up the form. The simulation results will be assessed by quantitative comparison with actual form development. From a computational point of view, as branching or organ forms are often represented by large and complex data-structures, we aim to develop optimized data structures and algorithms to achieve satisfactory compromises between accuracy and efficiency.

3.3. Axis3: Plasticity and robustness of forms

In this research axis, building on the insights gained from axes 1 and 2 on the mechanisms driving form development, we aim to explore the mechanistic origin of form plasticity and robustness. At the ontogenetic scale, we will study the ability of specific developmental mechanisms to buffer, or even to exploit, biological noise during morphogenesis. For plants, we will develop models capturing morphogenetic reactions to specific environmental changes (such as water stress or pruning), and their ability to modulate or even to reallocate growth in an opportunistic manner.

At the phylogenetic scale, we will investigate new connections that can be drawn from the use of a better understanding of form development mechanisms in the evolution of forms. In animals, we will use ascidians as a model organism to investigate how the variability of certain genomes relates to the variability of their forms. In plants, models of the genetic regulation of form development will be used to test hypotheses on the evolution of regulatory gene networks of key morphogenetic mechanisms such as branching. We believe that a better mechanistic understanding of developmental processes should shed new light on old evo-devo questions related to the evolution of biological forms, such as understanding the origin of *developmental constraints*[0] how the internal rules that govern form development, such as chemical interactions and physical constraints, may channel form changes so that selection is limited in the phenotype it can achieve?

3.4. Key modeling challenges

During the project lifetime, we will address several computational challenges related to the modeling of living forms and transversal to our main research axes. During the first phase of the project, we concentrate on 4 key challenges.

3.4.1. A new paradigm for modeling tree structures in biology

There is an ubiquitous presence of tree data in biology: plant structures, tree-like organs in animals (lungs, kidney vasculature), corals, sponges, but also phylogenic trees, cell lineage trees, *etc*. To represent, analyze and simulate these data, a huge variety of algorithms have been developed. For a majority, their computational time and space complexity is proportional to the size of the trees. In dealing with massive amounts of data, like trees in a plant orchard or cell lineages in tissues containing several thousands of cells, this level of complexity is often intractable. Here, our idea is to make use of a new class of tree structures, that can be efficiently compressed and that can be used to approximate any tree, to cut-down the complexity of usual algorithms on trees.

3.4.2. Efficient computational mechanical models of growing tissues

The ability to simulate efficiently physical forces that drive form development and their consequences in biological tissues is a critical issue of the MOSAIC project. Our aim is thus to design efficient algorithms to compute mechanical stresses within data-structures representing forms as the growth simulation proceeds. The challenge consists of computing the distribution of stresses and corresponding tissue deformations throughout data-structures containing thousands of 3D cells in close to interactive time. For this we will develop new strategies to simulate mechanics based on approaches originally developed in computer graphics to simulate in real time the deformation of natural objects. In particular, we will study how meshless and isogeometric variational methods can be adapted to the simulation of a population of growing and dividing cells.

3.4.3. Realistic integrated digital models

Most of the models developed in MOSAIC correspond to specific parts of real morphogenetic systems, avoiding the overwhelming complexity of real systems. However, as these models will be developed on computational structures representing the detailed geometry of an organ or an organism, it will be possible to assemble several of these sub-models within one single model, to figure out missing components, and to test potential interactions between the model sub-components as the form develops.

Throughout the project, we will thus develop two digital models, one plant and one animal, aimed at integrating various aspects of form development in a single simulation system. The development of these digital models will be made using an agile development strategy, in which the models are created and get functional at a very early stage, and become subsequently refined progressively.

[0] Raff, R. A. (1996). The Shape of Life: Genes, Development, and the Evolution of Form. Univ. Chicago Press.

3.4.4. Development of a computational environment for the simulation of biological form development

To support and integrate the software components of the team, we aim to develop a computational environment dedicated to the interactive simulation of biological form development. This environment will be built to support the paradigm of dynamical systems with dynamical structures. In brief, the form is represented at any time by a central data-structure that contains any topological, geometric, genetic and physiological information. The computational environment will provide in a user-friendly manner tools to up-load forms, to create them, to program their development, to analyze, visualize them and interact with them in 3D+time.

NEUROSYS Project-Team

3. Research Program

3.1. Main Objectives

The main challenge in computational neuroscience is the high complexity of neural systems. The brain is a complex system and exhibits a hierarchy of interacting subunits. On a specific hierarchical level, such subunits evolve on a certain temporal and spatial scale. The interactions of small units on a low hierarchical level build up larger units on a higher hierarchical level evolving on a slower time scale and larger spatial scale. By virtue of the different dynamics on each hierarchical level, until today the corresponding mathematical models and data analysis techniques on each level are still distinct. Only few analysis and modeling frameworks are known which link successfully at least two hierarchical levels.
After extracting models for different description levels, they are typically applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straightforward, it presents various difficulties. Usually the models involve a large set of unknown parameters which determine the dynamical properties of the models. To optimally reconstruct experimental features, it is necessary to formulate an inverse problem to extract optimally such model parameters from the experimental data. Typically this is a rather difficult problem due to the low signal-to-noise ratio in experimental brain signals. Moreover, the identification of signal features to be reconstructed by the model is not obvious in most applications. Consequently an extended analysis of the experimental data is necessary to identify the interesting data features. It is important to combine such a data analysis step with the parameter extraction procedure to achieve optimal results. Such a procedure depends on the properties of the experimental data and hence has to be developed for each application separately. Machine learning approaches that attempt to mimic the brain and its cognitive processes have had a lot of success in classification problems during the last decade. These hierarchical and iterative approaches use non-linear functions, which imitate neural cell responses, to communicate messages between neighboring layers. In our team, we work towards developing polysomnography-specific classifiers that might help in linking the features of particular interest for building systems for sleep signal classification with sleep mechanisms, with the accent on memory consolidation during the Rapid Eye Movement (REM) sleep phase.

3.2. Challenges

Techniques for the implementation and analysis of models achieved promises to be able to construct novel data monitors. This construction involves additional challenges and requires contact with realistic environments. By virtue of the specific applications of the research, close contact to hospitals and medical companies shall be established over a longer term in order to (i) gain deeper insight into the specific application of the devices and (ii) build specific devices in accordance with the actual need. Collaborations with local and national hospitals and the pharmaceutical industry already exist.

3.3. Research Directions

- From the microscopic to the mesoscopic scale:
  One research direction focuses on the *relation of single-neuron activity* on the microscopic scale *to the activity of neuronal populations*. To this end, the team investigates the stochastic dynamics of single neurons subject to external random inputs and involving random microscopic properties, such as random synaptic strengths and probability distributions of spatial locations of membrane ion channels. Such an approach yields a stochastic model of single neurons and allows the derivation of a stochastic neural population model.

  This bridge between the microscopic and mesoscopic scale may be performed via two pathways. The analytical and numerical treatment of the microscopic model may be called a *bottom-up approach*,

since it leads to a population activity model based on microscopic activity. This approach allows theoretical neural population activity to be compared to experimentally obtained population activity. The *top-down approach* aims at extracting signal features from experimental data gained from neural populations which give insight into the dynamics of neural populations and the underlying microscopic activity. The work on both approaches represents a well-balanced investigation of the neural system based on the systems properties.

- From the mesoscopic to the macroscopic scale:
  The other research direction aims to link neural population dynamics to macroscopic activity and behavior or, more generally, to phenomenological features. This link is more indirect but a very powerful approach to understand the brain, e.g., in the context of medical applications. Since real neural systems, such as in mammals, exhibit an interconnected network of neural populations, the team studies analytically and numerically the network dynamics of neural populations to gain deeper insight into possible phenomena, such as traveling waves or enhancement and diminution of certain neural rhythms. Electroencephalography (EEG) is a powerful brain imaging technique to study the overall brain activity in real time non-invasively. However it is necessary to develop robust techniques based on stable features by investigating the time and frequency domains of brain signals. Two types of information are typically used in EEG signals: (i) transient events such as evoked potentials, spindles and K-complexes and (ii) the power in specific frequency bands.

NUMED Project-Team

3. Research Program

3.1. Design of complex models

3.1.1. Project team positioning

The originality of our work is the quantitative description of phenomena accounting for several time and spatial scales. Here, propagation has to be understood in a broad sense. This includes propagation of invasive species, chemotactic waves of bacteira, evoluation of age structures populations ... Our main objectives are the quantitative calculation of macroscopic quantities as the rate of propagation, and microscopic distributions at the edge and the back of the front. These are essential features of propagation which are intimately linked in the long time dynamics.

3.1.2. Recent results

- Population models.

  H. Leman works at the interface between mathematics and biology, thanks to probabilist and determinist studies of models of populations. More precisely, she studies and develops probabilistic models, called agent models that described the population at an individual level. Each individual is characterized by one or more phenotypic traits and by its position, which may influence at the same time its ecological behavior and its motion. From a biological point of view these models are particularly interesting since they allow to include a large variety of interactions between individuals. These processes may also be studied in details to obtain theoretical results which may be simulated thanks to exact algorithms. To get quantitative results H. Leman uses changes of scales in space and time (large population, rare mutations, long time), following various biological assumptions.

  In a first study , H. Leman tries to understand the interactions between sexual preference mechanisms and evolutive forces inside spatially structured populations. Recently she got interesting in the description of necessary conditions to facilitate the emergence of such preferences by individuals.

  As a second example, H. Leman is also interested in the modeling and study of cooperative bacterias and tries to understand the impact of spatial structures in the eco - evolutions of these bacterias. Space seems to be an essential factor to facilitate the emergence of cooperation between bacterias.

  Finally, H. Leman studied the large time behavior of continuous state branching processes with competition and Lévy environment. These kind of stochastic processes are used to represent the fluctuations of the size of a population. In particular, she studied the extinction time of such a process.

- Inviscid limit of Navier Stokes equations.

  The question of the behavior of solutions of Navier Stokes equations in a bounded domain as the viscosity goes to 0 is a classical and highly difficult open question in Fluid Mechanics. A small boundary layer, called Prandtl layer, appears near the boundary, which turns out to be unstable if the viscosity is small enough. The stability analysis of this boundary layer is highly technical and remained open since the first formal analysis in the 1940's by physicists like Orr, Sommerfeld, Tollmien, Schlichting or Lin. E. Grenier recently made a complete mathematical analysis of this spectral problem, in collaboration with T. Nguyen and Y. Guo. We rigorously proved that any shear layer is spectrally and linearly unstable if the viscosity is small enough, which is the first mathematical result in that field. We also get some preliminary nonlinear results. A book on this subject is in preparation, already accepted by Springer.

- Numerical analysis of complex fluids: the example of avalanches.
  This deals with the development of numerical schemes for viscoplastic materials (namely with Bingham or Herschel-Bulkley laws). Recently, with other colleagues, Paul Vigneaux finished the design of the first 2D well-balanced finite volume scheme for a shallow viscoplastic model. It is illustrated on the famous Taconnaz avalanche path in the Mont-Blanc (see figure 1), Chamonix, in the case of dense snow avalanches. The scheme deals with general Digital Elevation Model (DEM) topographies, wet/dry fronts and is designed to compute precisely the stopping state of avalanches, a crucial point of viscoplastic flows which are able to rigidify [cf joint Figure and Fernandez-Nieto et al. JCP 2018]. Currently, through a collaboration with IRSTEA Grenoble, we also revisit the theory of viscoplastic boundary layers (see figure (2) by extending the Oldroyd's asymptotic scaling (1947) to the cases of moderate Bingham numbers (or Herschel-Bulkley numbers). Also with IRSTEA, we are developping a joint study (numerical and experimental) of viscoplastic avalanches in the lab, to challenge various yield stress models.

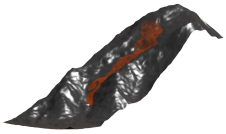


Figure 1. An example of avalanche simulation

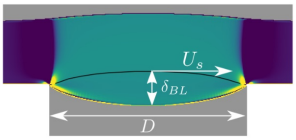


Figure 2. An example of boundary layer for complex flows

3.1.3. Collaborations

- Ecology: Orsay (C. Coron), Toulouse (IMT, M. Costa), MNHM Paris (V. Llaurens), LISC Paris (C. Smadi), ENS Paris (R. Ferrière, E. Abs), CIMAT (Mexique, J. C. P. Millan).
- Inviscid limit of Navier Stokes equations: Brown University (Y. Guo, B. Pausader), Penn State University (T. Nguyen), Orsay University (F. Rousset).
- Numerical analysis of complex fluids: Enrique D. Fernandez - Nieto (Univ. de Sevilla, Spain), Jose Maria Gallardo (Univ. de Malaga, Spain).
- Comparison between numerical simulations and physical experiments for the dam-break of viscoplastic materials: collaboration with IRSTEA (now INRAE, since Jan. 2020).

3.2. Parametrization of complex systems

3.2.1. Project-team positioning

Clinical data are often sparse: we have few data per patient. The number of data is of the order of the number of parameters. In this context, a natural way to parametrize complex models with real world clinical data is to use a Bayesian approach, namely to try to find the distribution of the model parameters in the population, rather than to try to identify the parameters of every single patient. This approach has been pioneered in the 90's by the Nonmem software, and has been much improved thanks to Marc Lavielle in the 2000's. Refined statistical methods, called SAEM, have been tuned and implemented in commercial softwares like Monolix.

3.2.2. Recent results

The main problem when we try to parametrize clinical data using complex systems is the computational time. One single evaluation of the model can be costly, in particular if this model involves partial differential equations, and SAEM algorithm requires hundreds of thousands of single evaluations. The time cost is then too large, in particular because SAEM may not be parallelized.

To speed up the evaluation of the complex model, we replace it by an approximate one, or so called metamodel, constructed by interpolation of a small number of its values. We therefore combine the classical SAEM algorithm with an interpolation step, leading to a strong acceleration. Interpolation can be done through a precomputation step on a fixed grid, or through a more efficient kriging step. The interpolation grid or the kriging step may be improved during SAEM algorithm in an iterative way in order to get accurate evaluations of the complex system only in the domain of interest, namely near the clinical values [14],[15].

We applied these new algorithms to synthetic data and are currently using them on glioma data. We are also currently trying to prove the convergence of the corresponding algorithms. We will develop glioma applications in the next section.

Moreover E. Ollier in his phD developed new strategies to distinguish various populations within a SAEM algorithm [23].

We have two long standing collaborations with Sanofi and Servier on parametrization issues:

- Servier: during a four years contract, we modelled the pkpd of new drugs and also study the combination and optimization of chimiotherapies.
- Sanofi: during a eight years contract, Emmanuel Grenier wrote a complete software devoted to the study of the degradation of vaccine. This software is used worldwide by Sanofi R&D teams in order to investigate the degradation of existing or new vaccines and to study their behavior when they are heated. This software has been used on flu, dengue and various other diseases.

3.2.3. Collaborations

- Academic collaborations: A. Leclerc Samson (Grenoble University)
- Medical collaborations: Dr Ducray (Centre Léon Bérard, Lyon) and Dr Sujobert (Lyon Sud Hospital)
- Industrial contracts: we used parametrization and treatment improvement techniques for Servier (four years contract, on cancer drug modeling and optimization) and Sanofi (long standing collaboration)

3.3. Multiscale models in oncology

3.3.1. Project-team positioning

Cancer modeling is the major topic of several teams in France and Europe, including Mamba, Monc and Asclepios to quote only a few Inria teams. These teams try to model metastasis, tumoral growth, vascularisation through angiogenesis, or to improve medical images quality. Their approaches are based on dynamical systems, partial differential equations, or on special imagery techniques.

Numed focuses on the link between very simple partial differential equations models, like reaction diffusion models, and clinical data.

3.3.2. Results

During 2018 we developed new collaborations with the Centre Léon Bérard (Lyon), in particular on the following topics

- Barcoding of cells: thanks to recent techniques, it is possible to mark each cell with an individual barcode, and to follow its division and descendance. The analysis of such data requires probabilistic models, in particular to model experimental bias.

- Apoptosis: the question is to investigate whether the fate of neighboring cells influence the evolution of a given cell towards apoptosis, starting from videos of in vitro drug induced apoptosis.
- Dormance: Study of the dynamics of cells under immunotherapy, starting from experimental in vitro data.
- Colorectal cancer: In vitro study of the role of stem cells in drug resistance, in colorectal cancer.

3.3.3. Collaborations

- Centre Léon Bérard (in particular: Pr Puisieux, G. Ichim, M. Plateroni, S. Ortiz).

OPIS Project-Team

3. Research Program

3.1. Accelerated algorithms for solving high-dimensional continuous optimization problems

Variational problems requiring the estimation of a huge number of variables have now to be tackled, especially in the field of 3D reconstruction/restoration (e.g. $\geq 10^9$ variables in 3D imaging). In addition to the curse of dimensionality, another difficulty to overcome is that the cost function usually reads as the sum of several loss/regularization terms, possibly composed with large-size linear operators. These terms can be nonsmooth and/or nonconvex, as they may serve to promote the sparsity of the sought solution in some suitable representation (e.g. a frame) or to fufill some physical constraints. In such a challenging context, there is a strong need for developing fast parallelized optimization agorithms for which sound theoretical guarantees of convergence can be established. We explore deterministic and stochastic approaches based on proximal tools, MM (Majorization-Minimization) strategies, and trust region methods. Because of the versatility of the methods that will be proposed, a wide range of applications in image recovery are considered: parallel MRI, breast tomosynthesis, 3D ultrasound imaging, and two-photon microscopy. For example, in breast tomosynthesis (collaboration with GE Healthcare), 3D breast images have to be reconstructed from a small number of X-ray projections with limited view angles. Our objective is to facilitate the clinical task by developing advanced reconstruction methods allowing micro-calcifications to be highlighted. In two-photon microscopy (collaboration with XLIM), our objective is to provide effective numerical solutions to improve the 3D resolution of the microscope, especially when cheap laser sources are used, with applications to muscle disease screening.

3.2. Optimization over graphs

Graphs and hypergraphs are rich data structures for capturing complex, possibly irregular, dependencies in multidimensional data. Coupled with Markov models, they constitute the backbones of many techniques used in computer vision. Optimization is omnipresent in graph processing. Firstly, it allows the structure of the underlying graph to be inferred from the observed data, when the former is hidden. Second, it permits to develop graphical models based on the prior definition of a meaningful cost function. This leads to powerful nonlinear estimates of variables corresponding to unknown weights on the vertices and/or the edges of the graph. Tasks such as partitioning the graph into subgraphs corresponding to different clusters (e.g., communities in social networks) or graph matching, can effectively be performed within this framework. Finally, graphs by themselves offer flexible structures for formulating and solving optimization problems in an efficient distributed manner. On all these topics, our group has acquired a long-term expertise that we plan to further strengthen. In terms of applications, novel graph mining methods are proposed for gene regulatory and brain network analysis. For example, we plan to develop sophisticated methods for better understanding the gene regulatory network of various microscopic fungi, in order to improve the efficiency of the production of bio-fuels (collaboration with IFP Energies Nouvelles).

3.3. Toward more understandable deep learning

Nowadays, deep learning techniques efficiently solve supervised tasks in classification or regression by utilizing large amounts of labeled data and the powerful high level features that they learn by using the input data. Their good performance has caught the attention of the optimization community since currently these methods offer virtually no guarantee of convergence, stability or generalization. Deep neural networks are optimized through a computationally intensive engineering process via methods based on stochastic gradient descent. These methods are slow and they may not lead to relevant local minima. Thus, more efforts must

be dedicated in order to improve the training of deep neural networks by proposing better optimization algorithms applicable to large-scale datasets. Beyond optimization, incorporating some structure in deep neural networks permits more advanced regularization than the current methods. This should reduce their complexity, as well as allow us to derive some bounds regarding generalization. For example, many signal processing models (e.g. those based on multiscale decompositions) exhibit some strong correspondence with deep learning architectures, yet they do not require as many parameters. One can thus think of introducing some supervision into these models in order to improve their performance on standard benchmarks. A better mathematical understanding of these methods permits to improve them, but also to propose some new models and representations for high-dimensional data. This is particularly interesting in settings such as the diagnosis or prevention of diseases from medical images, because they correspond to critical applications where the made decision is crucial and needs to be interpretable. One of the main applications of this work is to propose robust models for the prediction of the outcome of cancer immunotherapy treatments from multiple and complementary sources of information: images, gene expression data, patient profile, etc (collaboration with Institut Gustave Roussy).

PARIETAL Project-Team

3. Research Program

3.1. Inverse problems in Neuroimaging

Many problems in neuroimaging can be framed as forward and inverse problems. For instance, brain population imaging is concerned with the *inverse problem* that consists in predicting individual information (behavior, phenotype) from neuroimaging data, while the corresponding *forward problem* boils down to explaining neuroimaging data with the behavioral variables. Solving these problems entails the definition of two terms: a loss that quantifies the goodness of fit of the solution (does the model explain the data well enough?), and a regularization scheme that represents a prior on the expected solution of the problem. These priors can be used to enforce some properties on the solutions, such as sparsity, smoothness or being piece-wise constant.

Let us detail the model used in typical inverse problem: Let $\mathbf{X}$ be a neuroimaging dataset as an $(n_{subjects}, n_{voxels})$ matrix, where $n_{subjects}$ and n_{voxels} are the number of subjects under study, and the image size respectively, $\mathbf{Y}$ a set of values that represent characteristics of interest in the observed population, written as $(n_{subjects}, n_{features})$ matrix, where $n_{features}$ is the number of characteristics that are tested, and $\mathbf{w}$ an array of shape $(n_{voxels}, n_{features})$ that represents a set of pattern-specific maps. In the first place, we may consider the columns $\mathbf{Y}_1, .., \mathbf{Y}_{n_{features}}$ of Y independently, yielding $n_{features}$ problems to be solved in parallel:

$$\mathbf{Y}_i = \mathbf{X}\mathbf{w}_i + \epsilon_i, \forall i \in \{1, .., n_{features}\},$$

where the vector contains $\mathbf{w}_i$ is the i^{th} row of $\mathbf{w}$. As the problem is clearly ill-posed, it is naturally handled in a regularized regression framework:

$$\widehat{w}_i = \text{argmin}_{w_i} \|\mathbf{Y}_i - \mathbf{X}\mathbf{w}_i\|^2 + \Psi(\mathbf{w}_i), \tag{59}$$

where Ψ is an adequate penalization used to regularize the solution:

$$\Psi(\mathbf{w}; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2 + \eta_1 \|\nabla \mathbf{w}\|_{2,1} + \eta_2 \|\nabla \mathbf{w}\|_{2,2} \tag{60}$$

with λ_1, λ_2, η_1, $\eta_2 \geq 0$ (this formulation particularly highlights the fact that convex regularizers are norms or quasi-norms). In general, only one or two of these constraints is considered (hence is enforced with a non-zero coefficient):

- When $\lambda_1 > 0$ only (LASSO), and to some extent, when $\lambda_1, \lambda_2 > 0$ only (elastic net), the optimal solution $\mathbf{w}$ is (possibly very) sparse, but may not exhibit a proper image structure; it does not fit well with the intuitive concept of a brain map.
- Total Variation regularization (see Fig. 1) is obtained for ($\eta_1 > 0$ only), and typically yields a piece-wise constant solution. It can be associated with Lasso to enforce both sparsity and sparse variations.
- Smooth lasso is obtained with ($\eta_2 > 0$ and $\lambda_1 > 0$ only), and yields smooth, compactly supported spatial basis functions.

Note that, while the qualitative aspect of the solutions are very different, the predictive power of these models is often very close.

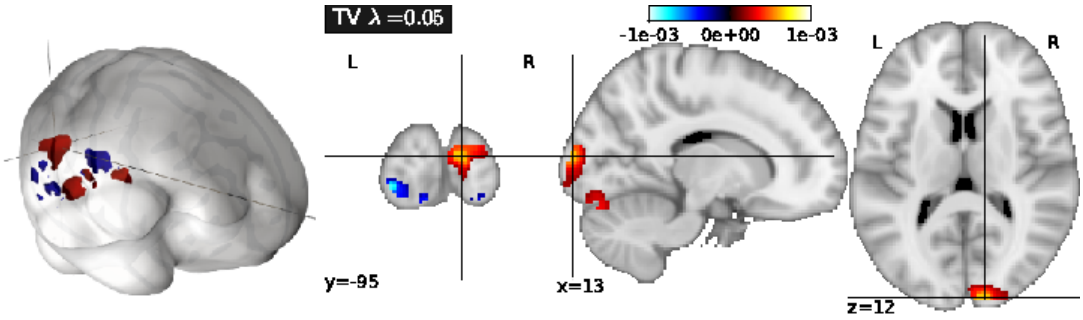


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here is to predict the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the presentation of an image. Learning and test are performed across individuals. Unlike other approaches, Total Variation regularization yields a sparse and well-localized solution that also enjoys high predictive accuracy.

The performance of the predictive model can simply be evaluated as the amount of variance in $\mathbf{Y}_i$ fitted by the model, for each $i \in \{1, .., n_{features}\}$. This can be computed through cross-validation, by *learning* $\widehat{\mathbf{w}}_i$ on some part of the dataset, and then estimating $\|\mathbf{Y}_i - \mathbf{X}\widehat{w}_i\|^2$ using the remainder of the dataset.

This framework is easily extended by considering

- *Grouped penalization*, where the penalization explicitly includes a prior clustering of the features, i.e. voxel-related signals, into given groups. This amounts to enforcing structured priors on the solution.
- *Combined penalizations*, i.e. a mixture of simple and group-wise penalizations, that allow some variability to fit the data in different populations of subjects, while keeping some common constraints.
- *Logistic and hinge regression*, where a non-linearity is applied to the linear model so that it yields a probability of classification in a binary classification problem.
- *Robustness to between-subject variability* to avoid the learned model overly reflecting a few outlying particular observations of the training set. Note that noise and deviating assumptions can be present in both $\mathbf{Y}$ and $\mathbf{X}$
- *Multi-task learning*: if several target variables are thought to be related, it might be useful to constrain the estimated parameter vector $\mathbf{w}$ to have a shared support across all these variables.
  For instance, when one of the variables $\mathbf{Y}_i$ is not well fitted by the model, the estimation of other variables $\mathbf{Y}_j, j \neq i$ may provide constraints on the support of $\mathbf{w}_i$ and thus, improve the prediction of $\mathbf{Y}_i$.

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \epsilon, \tag{61}$$

then

$$\widehat{w} = \text{argmin}_{\mathbf{w}=(\mathbf{w}_i), i=1..n_f} \sum_{i=1}^{n_f} \|\mathbf{Y_i} - \mathbf{X}\mathbf{w_i}\|^2 + \lambda \sum_{j=1}^{n_{voxels}} \sqrt{\sum_{i=1}^{n_f} \mathbf{w_{i,j}^2}} \tag{62}$$

3.2. Multivariate decompositions

Multivariate decompositions provide a way to model complex data such as brain activation images: for instance, one might be interested in extracting an *atlas of brain regions* from a given dataset, such as regions exhibiting similar activity during a protocol, across multiple protocols, or even in the absence of protocol (during resting-state). These data can often be factorized into spatial-temporal components, and thus can be estimated through *regularized Principal Components Analysis* (PCA) algorithms, which share some common steps with regularized regression.
Let $\mathbf{X}$ be a neuroimaging dataset written as an $(n_{subjects}, n_{voxels})$ matrix, after proper centering; the model reads

$$\mathbf{X} = \mathbf{AD} + \epsilon, \tag{63}$$

where $\mathbf{D}$ represents a set of n_{comp} spatial maps, hence a matrix of shape (n_{comp}, n_{voxels}), and $\mathbf{A}$ the associated subject-wise loadings. While traditional PCA and independent components analysis (ICA) are limited to reconstructing components $\mathbf{D}$ within the space spanned by the column of $\mathbf{X}$, it seems desirable to add some constraints on the rows of $\mathbf{D}$, that represent spatial maps, such as sparsity, and/or smoothness, as it makes the interpretation of these maps clearer in the context of neuroimaging. This yields the following estimation problem:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{AD}\|^2 + \Psi(\mathbf{D}) \text{ s.t. } \|\mathbf{A}_i\| = 1 \; \forall i \in \{1..n_{features}\}, \tag{64}$$

where $(\mathbf{A}_i),\ i \in \{1..n_{features}\}$ represents the columns of $\mathbf{A}$. Ψ can be chosen such as in Eq. (2) in order to enforce smoothness and/or sparsity constraints.
The problem is not jointly convex in all the variables but each penalization given in Eq (2) yields a convex problem on $\mathbf{D}$ for $\mathbf{A}$ fixed, and conversely. This readily suggests an alternate optimization scheme, where $\mathbf{D}$ and $\mathbf{A}$ are estimated in turn, until convergence to a local optimum of the criterion. As in PCA, the extracted components can be ranked according to the amount of fitted variance. Importantly, also, estimated PCA models can be interpreted as a probabilistic model of the data, assuming a high-dimensional Gaussian distribution (probabilistic PCA).

Ultimately, the main limitations to these algorithms is the cost due to the memory requirements: holding datasets with large dimension and large number of samples (as in recent neuroimaging cohorts) leads to inefficient computation. To solve this issue, online methods are particularly attractive [1].

3.3. Covariance estimation

Another important estimation problem stems from the general issue of learning the relationship between sets of variables, in particular their covariance. Covariance learning is essential to model the dependence of these variables when they are used in a multivariate model, for instance to study potential interactions among them and with other variables. Covariance learning is necessary to model latent interactions in high-dimensional observation spaces, e.g. when considering multiple contrasts or functional connectivity data.
The difficulties are two-fold: on the one hand, there is a shortage of data to learn a good covariance model from an individual subject, and on the other hand, subject-to-subject variability poses a serious challenge to the use of multi-subject data. While the covariance structure may vary from population to population, or depending on the input data (activation versus spontaneous activity), assuming some shared structure across problems, such as their sparsity pattern, is important in order to obtain correct estimates from noisy data. Some of the most important models are:

- **Sparse Gaussian graphical models**, as they express meaningful conditional independence relationships between regions, and do improve conditioning/avoid overfit.

- **Decomposable models**, as they enjoy good computational properties and enable intuitive interpretations of the network structure. Whether they can faithfully or not represent brain networks is still an open question.
- **PCA-based regularization of covariance** which is powerful when modes of variation are more important than conditional independence relationships.

Adequate model selection procedures are necessary to achieve the right level of sparsity or regularization in covariance estimation; the natural evaluation metric here is the out-of-sample likelihood of the associated Gaussian model. Another essential remaining issue is to develop an adequate statistical framework to test differences between covariance models in different populations. To do so, we consider different means of parametrizing covariance distributions and how these parametrizations impact the test of statistical differences across individuals.

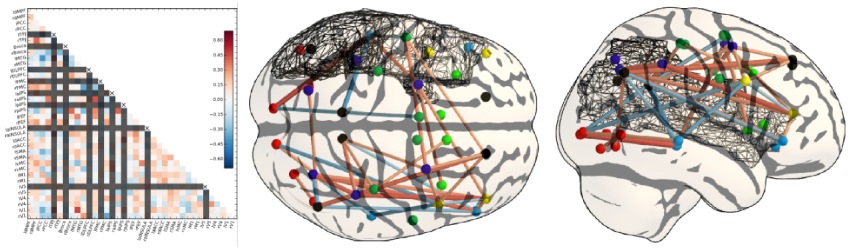


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical detection of the difference requires a sophisticated statistical framework for the comparison of graphical models.

PLEIADE Project-Team

3. Research Program

3.1. A Geometric View of Diversity

Diversity may be studied as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, it is possible, after computation of pairwise distances, or sometimes dissimilarities, to build a Euclidean image of it as a point cloud in a space of relevant dimension. Then, diversity can be associated with the shape of the point cloud. The human eye is often far better than an algorithm at recognizing a pattern or shape. One objective of our project is to narrow the gap between the story that a human eye can tell, and that an algorithm can tell. Several directions will be explored. First, this requires mastering classical tools in dimension reduction, mainly algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...). Second, neighborhoods in point clouds naturally lead to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points (representing, say, DNA sequences) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemannian geometry). It is expected that some properties of the manifold can tell something of the constraints on the space where measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as mesh embedded in a manifold, is currently an active field of research [28], [27]. See as well [30] for a link between geometric structure, linear and nonlinear dimensionality reduction.

Biodiversity and high-performance computing: Most methods and tools for characterizing diversity have been designed for datasets that can be analyzed on a laptop, but NGS datasets produced for metabarcoding are far too large. Data analysis algorithms and tools must be revisited and scaled up. We will mobilize both distributed algorithms like the Arnoldi method and new algorithms, like random projection or column selection methods, to build point clouds in Euclidean spaces from massive data sets, and thus to overcome the cubic complexity of computation of eigenvectors and eigenvalues of very large dense matrices. We will also link distance geometry [22] with convex optimization procedures through matrix completion [15], [17].

Intercalibration: There is a considerable difference between supervised and unsupervised clustering: in supervised clustering, the result for an item i is independent from the result for an item $j \neq i$, whereas in unsupervised clustering, the result for an item i (e.g. the cluster it belongs to, and its composition) depends on nearby items $j \neq i$. Which means that the result may change if some items are added to or subtracted from the sample. This raises the more global problem of how to merge two studies to yield a more comprehensive view of biodiversity?

3.2. Knowledge Management for Biology

The heterogenous data generated in computational molecular biology and ecology are distinguished not only by their volume, but by the richness of the many levels of interpretation that biologists create. The same nucleic acid sequence can be seen as a molecule with a structure, a sequence of base pairs, a collection of genes, an allele, or a molecular fingerprint. To extract the maximum benefit from this treasure trove we must organize the knowledge in ways that facilitate extraction, analysis, and inference. Our focus has been on the efficient representation of relations between biological objects and operations on those representations, in particular heuristic analyses and logical inference.

PLEIADE will develop applications in comparative genomics of related organisms, using new mathematical tools for representing compactly, at different scales of difference, comparisons between related genomes. New methods based on distance geometry will refine these comparisons. Compact representations can be stored, exchanged, and combined. They will form the basis of new simultaneous genome annotation methods, linked directly to abductive inference methods for building functional models of the organisms and their communities.

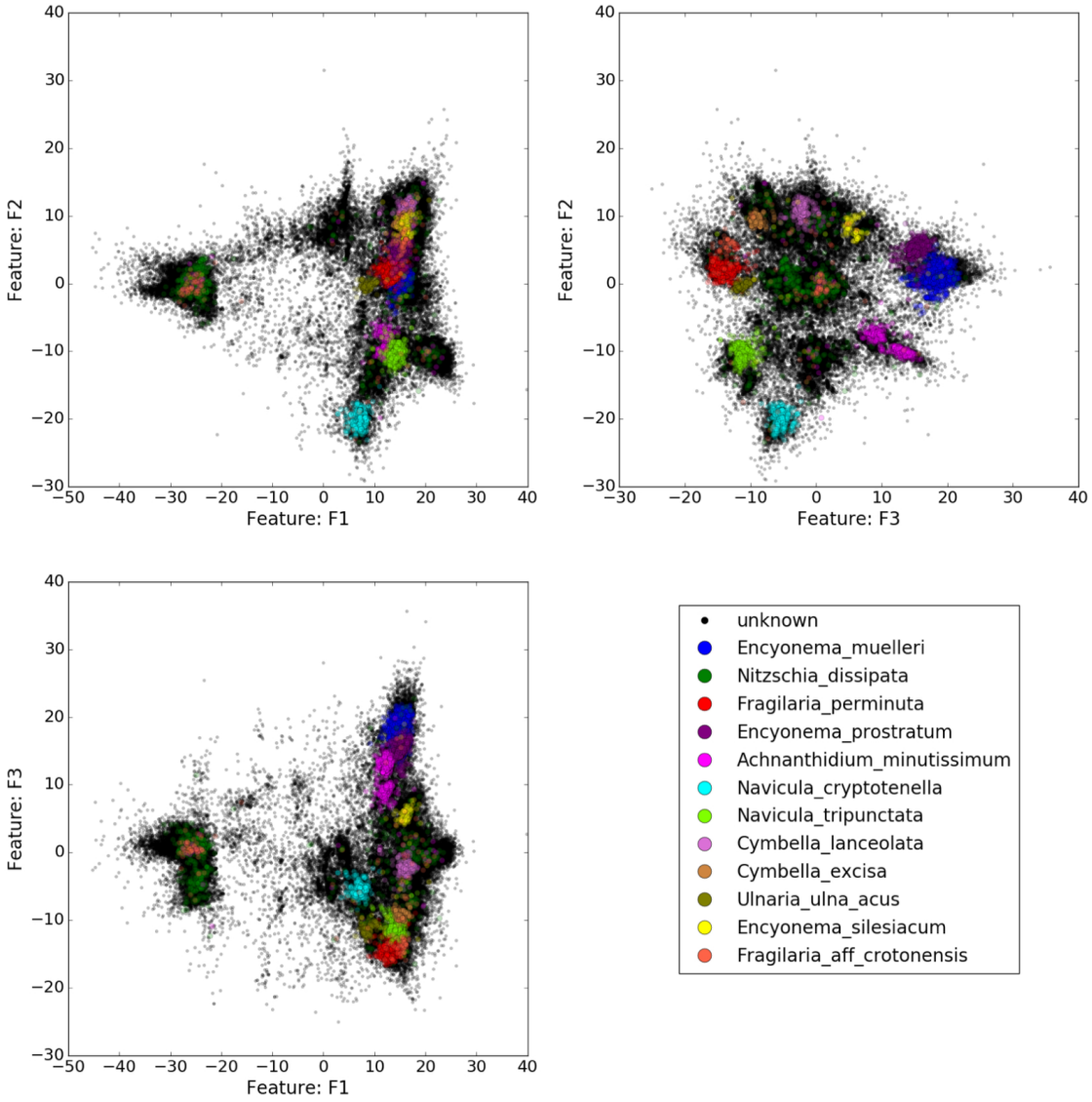


Figure 3. Validation of high density islands using supervised classification. Metagenomic reads from diatoms in Lake Geneva [26] were analyzed by the method from [16] and colored by species according to a reference database.

Since a goal of PLEIADE is to integrate diversity throughout the analysis process, it is necessary to incorporate **diversity as a form of knowledge** that can be stored in a knowledge base. Diversity can be represented using various compact representations, such as trees and quotient graphs storing nested sets of relations. Extracting structured representations and logical relations from integrated knowledge bases (Figure 2) will require domain-specific query methods that can express forms of diversity.

3.3. Modeling by successive refinement

Describing the links between diversity in traits and diversity in function will require comprehensive models, assembled from and refining existing models. A recurring difficulty in building comprehensive models of biological systems is that accurate models for subsystems are built using different formalisms and simulation techniques, and hand-tuned models tend to be so focused in scope that it is difficult to repurpose them [13]. Our belief is that a sustainable effort in building efficient behavioral models must proceed incrementally, rather than by modeling individual processes *de novo*. *Hierarchical modeling* [10] is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have previously shown that this approach can be effective for certains kinds of systems in biotechnology [2], [14] and medicine [12]. Our challenge is to adapt incremental, hierarchical refinement to modeling organisms and communities in metagenomic and comparative genomic applications.

REO Team

3. Research Program

3.1. Multiphysics modeling

In large vessels and in large bronchi, blood and air flows are generally supposed to be governed by the incompressible Navier-Stokes equations. Indeed in large arteries, blood can be supposed to be Newtonian, and at rest air can be modeled as an incompressible fluid. The cornerstone of the simulations is therefore a Navier-Stokes solver. But other physical features have also to be taken into account in simulations of biological flows, in particular fluid-structure interaction in large vessels and transport of sprays, particles or chemical species.

3.1.1. Fluid-structure interaction

Fluid-structure coupling occurs both in the respiratory and in the circulatory systems. We focus mainly on blood flows since our work is more advanced in this field. But the methods developed for blood flows could be also applied to the respiratory system.

Here "fluid-structure interaction" means a coupling between the 3D Navier-Stokes equations and a 3D (possibly thin) structure in large displacements.

The numerical simulations of the interaction between the artery wall and the blood flows raise many issues: (1) the displacement of the wall cannot be supposed to be infinitesimal, geometrical nonlinearities are therefore present in the structure and the fluid problem have to be solved on a moving domain (2) the densities of the artery walls and the blood being close, the coupling is strong and has to be tackled very carefully to avoid numerical instabilities, (3) "naive" boundary conditions on the artificial boundaries induce spurious reflection phenomena.

Simulation of valves, either at the outflow of the cardiac chambers or in veins, is another example of difficult fluid-structure problems arising in blood flows. In addition, very large displacements and changes of topology (contact problems) have to be handled in those cases.

Due to stability reasons, it seems impossible to successfully apply in hemodynamics the explicit coupling schemes used in other fluid-structure problems, like aeroelasticity. As a result, fluid-structure interaction in biological flows raise new challenging issues in scientific computing and numerical analysis : new schemes have to be developed and analyzed.

We have proposed and analyzed over the last few years several efficient fluid-structure interaction algorithms. This topic remains very active. We are now using these algorithms to address inverse problems in blood flows to make patient specific simulations (for example, estimation of artery wall stiffness from medical imaging).

3.1.2. Aerosol

Complex two-phase fluids can be modeled in many different ways. Eulerian models describe both phases by physical quantities such as the density, velocity or energy of each phase. In the mixed fluid-kinetic models, the biphasic fluid has one dispersed phase, which is constituted by a spray of droplets, with a possibly variable size, and a continuous classical fluid.

This type of model was first introduced by Williams [36] in the frame of combustion. It was later used to develop the Kiva code [26] at the Los Alamos National Laboratory, or the Hesione code [31], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [29], therapeutic sprays, *etc.* One of the interests of such a model is that various phenomena on the droplets can be taken into account with an accurate precision: collision, breakups, coagulation, vaporization, chemical reactions, *etc.*, at the level of the droplets.

The model usually consists in coupling a kinetic equation, that describes the spray through a probability density function, and classical fluid equations (typically Navier-Stokes). The numerical solution of this system relies on the coupling of a method for the fluid equations (for instance, a finite volume method) with a method fitted to the spray (particle method, Monte Carlo).

We are mainly interested in modeling therapeutic sprays either for local or general treatments. The study of the underlying kinetic equations should lead us to a global model of the ambient fluid and the droplets, with some mathematical significance. Well-chosen numerical methods can give some tracks on the solutions behavior and help to fit the physical parameters which appear in the models.

3.2. Multiscale modeling

Multiscale modeling is a necessary step for blood and respiratory flows. In this section, we focus on blood flows. Nevertheless, similar investigations are currently carried out on respiratory flows.

3.2.1. Arterial tree modeling

Problems arising in the numerical modeling of the human cardiovascular system often require an accurate description of the flow in a specific sensible subregion (carotid bifurcation, stented artery, *etc.*). The description of such local phenomena is better addressed by means of three-dimensional (3D) simulations, based on the numerical approximation of the incompressible Navier-Stokes equations, possibly accounting for compliant (moving) boundaries. These simulations require the specification of boundary data on artificial boundaries that have to be introduced to delimit the vascular district under study. The definition of such boundary conditions is critical and, in fact, influenced by the global systemic dynamics. Whenever the boundary data is not available from accurate measurements, a proper boundary condition requires a mathematical description of the action of the reminder of the circulatory system on the local district. From the computational point of view, it is not affordable to describe the whole circulatory system keeping the same level of detail. Therefore, this mathematical description relies on simpler models, leading to the concept of *geometrical multiscale* modeling of the circulation [32]. The underlying idea consists in coupling different models (3D, 1D or 0D) with a decreasing level of accuracy, which is compensated by their decreasing level of computational complexity.

The research on this topic aims at providing a correct methodology and a mathematical and numerical framework for the simulation of blood flow in the whole cardiovascular system by means of a geometric multiscale approach. In particular, one of the main issues will be the definition of stable coupling strategies between 3D and reduced order models.

To model the arterial tree, a standard way consists of imposing a pressure or a flow rate at the inlet of the aorta, *i.e.* at the network entry. This strategy does not allow to describe important features as the overload in the heart caused by backward traveling waves. Indeed imposing a boundary condition at the beginning of the aorta artificially disturbs physiological pressure waves going from the arterial tree to the heart. The only way to catch this physiological behavior is to couple the arteries with a model of heart, or at least a model of left ventricle.

A constitutive law for the myocardium, controlled by an electrical command, has been developed in the CardioSense3D project [0]. One of our objectives is to couple artery models with this heart model.

A long term goal is to achieve 3D simulations of a system including heart and arteries. One of the difficulties of this very challenging task is to model the cardiac valves. To this purpose, we investigate a mix of arbitrary Lagrangian Eulerian and fictitious domain approaches or x-fem strategies, or simplified valve models based on an immersed surface strategy.

[0] http://www-sop.inria.fr/CardioSense3D/

3.2.2. Heart perfusion modeling

The heart is the organ that regulates, through its periodical contraction, the distribution of oxygenated blood in human vessels in order to nourish the different parts of the body. The heart needs its own supply of blood to work. The coronary arteries are the vessels that accomplish this task. The phenomenon by which blood reaches myocardial heart tissue starting from the blood vessels is called in medicine perfusion. The analysis of heart perfusion is an interesting and challenging problem. Our aim is to perform a three-dimensional dynamical numerical simulation of perfusion in the beating heart, in order to better understand the phenomena linked to perfusion. In particular the role of the ventricle contraction on the perfusion of the heart is investigated as well as the influence of blood on the solid mechanics of the ventricle. Heart perfusion in fact implies the interaction between heart muscle and blood vessels, in a sponge-like material that contracts at every heartbeat via the myocardium fibers.

Despite recent advances on the anatomical description and measurements of the coronary tree and on the corresponding physiological, physical and numerical modeling aspects, the complete modeling and simulation of blood flows inside the large and the many small vessels feeding the heart is still out of reach. Therefore, in order to model blood perfusion in the cardiac tissue, we must limit the description of the detailed flows at a given space scale, and simplify the modeling of the smaller scale flows by aggregating these phenomena into macroscopic quantities, by some kind of "homogenization" procedure. To that purpose, the modeling of the fluid-solid coupling within the framework of porous media appears appropriate.

Poromechanics is a simplified mixture theory where a complex fluid-structure interaction problem is replaced by a superposition of both components, each of them representing a fraction of the complete material at every point. It originally emerged in soils mechanics with the work of Terzaghi [35], and Biot [27] later gave a description of the mechanical behavior of a porous medium using an elastic formulation for the solid matrix, and Darcy's law for the fluid flow through the matrix. Finite strain poroelastic models have been proposed (see references in [28]), albeit with *ad hoc* formulations for which compatibility with thermodynamics laws and incompressibility conditions is not established.

3.2.3. Tumor and vascularization

The same way the myocardium needs to be perfused for the heart to beat, when it has reached a certain size, tumor tissue needs to be perfused by enough blood to grow. It thus triggers the creation of new blood vessels (angiogenesis) to continue to grow. The interaction of tumor and its micro-environment is an active field of research. One of the challenges is that phenomena (tumor cell proliferation and death, blood vessel adaptation, nutrient transport and diffusion, etc) occur at different scales. A multi-scale approach is thus being developed to tackle this issue. The long term objective is to predict the efficiency of drugs and optimize therapy of cancer.

3.2.4. Respiratory tract modeling

We aim at developing a multiscale model of the respiratory tract. Intraprenchymal airways distal from generation 7 of the tracheabronchial tree (TBT), which cannot be visualized by common medical imaging techniques, are modeled either by a single simple model or by a model set according to their order in TBT. The single model is based on straight pipe fully developed flow (Poiseuille flow in steady regimes) with given alveolar pressure at the end of each compartment. It will provide boundary conditions at the bronchial ends of 3D TBT reconstructed from imaging data. The model set includes three serial models. The generation down to the pulmonary lobule will be modeled by reduced basis elements. The lobular airways will be represented by a fractal homogenization approach. The alveoli, which are the gas exchange loci between blood and inhaled air, inflating during inspiration and deflating during expiration, will be described by multiphysics homogenization.

SERENA Project-Team

3. Research Program

3.1. Multiphysics coupling

Within our project, we start from the conception and analysis of *models* based on *partial differential equations* (PDEs). Already at the PDE level, we address the question of *coupling* of different models; examples are that of simultaneous fluid flow in a discrete network of two-dimensional *fractures* and in the surrounding three-dimensional porous medium, or that of interaction of a compressible flow with the surrounding elastic *deformable structure*. The key physical characteristics need to be captured, whereas existence, uniqueness, and continuous dependence on the data are minimal analytic requirements that we seek to satisfy. At the modeling stage, we also develop model-order reduction techniques, such as the use of reduced basis techniques or proper generalized decompositions, to tackle evolutive problems, in particular in the nonlinear case, and we are also interested in developing reduced-order methods for variational inequalities such as those encountered in solid mechanics with contact and possibly also friction.

3.2. Discretization by hybrid high-order and discrete element methods

We consequently design *numerical methods* for the devised model. Traditionally, we have worked in the context of finite element, finite volume, mixed finite element, and discontinuous Galerkin methods. Novel classes of schemes enable the use of general *polygonal* and *polyhedral meshes* with *nonmatching interfaces*, and we develop them in response to a high demand from our industrial partners (namely EDF, CEA, and IFP Energies Nouvelles). In the lowest-order case, our focus is to design *discrete element* methods for solid mechanics. The novelty is to devise these methods to treat dynamic elastoplasticity as well as quasi-static and dynamic crack propagation. We also develop *structure-preserving* methods for the Navier–Stokes equations, i.e., methods that mimic algebraically at the discrete level fundamental properties of the underlying PDEs, such as conservation principles and preservation of invariants. In the higher-order case, we actively contribute to the development of *hybrid high-order* methods. We contribute to the numerical analysis in nonlinear cases (obstacle problem, Signorini conditions), we apply these methods to challenging problems from solid mechanics involving large deformations and plasticity, and we develop a comprehensive software implementing them. We believe that these methods belong to the future generation of numerical methods for industrial simulations; as a concrete example, the implementation of these methods in an industrial software of EDF has been completed in 2019 in the framework of the PhD thesis of Nicolas Pignet.

3.3. Domain decomposition and Newton–Krylov (multigrid) solvers

We next concentrate an intensive effort on the development and analysis of efficient solvers for the systems of nonlinear algebraic equations that result from the above discretizations. We have in the past developed *Newton–Krylov solvers* like the *adaptive inexact Newton method*, and we place a particular emphasis on *parallelization* achieved via the *domain decomposition* method. Here we traditionally specialize in *Robin transmission conditions*, where an optimized choice of the parameter has already shown speed-ups in orders of magnitude in terms of the number of domain decomposition iterations in model cases. We concentrate in the SERENA project on adaptation of these algorithms to the above novel discretization schemes, on the optimization of the free Robin parameter for challenging situations, and also on the use of the Ventcell transmission conditions. Another feature is the use of such algorithms in time-dependent problems in *space-time* domain decomposition that we have recently pioneered. This allows the use of different time steps in different parts of the computational domain and turns out to be particularly useful in porous media applications, where the amount of diffusion (permeability) varies abruptly, so that the evolution speed varies significantly from one part of the computational domain to another. Our new theme here are *Newton–multigrid solvers*, where the geometric multigrid solver is *tailored* to the specific problem under consideration and to the specific

numerical method, with problem- and discretization-dependent restriction, prolongation, and smoothing. Using patchwise smoothing, we have in particular recently developed a first multigrid method whose behavior is both in theory and in practice insensitive of (robust with respect to) the approximation polynomial degree. With patchwise techniques, we also achieve mass balance at each iteration step, a highly demanded feature in most of the target applications. The solver itself is then *adaptively steered* at each execution step by an a posteriori error estimate (adaptive stepsize, adaptive smoothing).

3.4. Reliability by a posteriori error control

The fourth part of our theoretical efforts goes towards guaranteeing the results obtained at the end of the numerical simulation. Here a key ingredient is the development of rigorous *a posteriori estimates* that make it possible to estimate in a fully computable way the error between the unknown exact solution and its numerical approximation. Our estimates also allow to distinguish the different *components* of the overall *error*, namely the errors coming from modeling, from the discretization scheme, from the nonlinear (Newton) solver, and from the linear algebraic (Krylov, domain decomposition, multigrid) solver. A new concept here is that of *local stopping criteria*, where all the error components are balanced locally within each computational mesh element. This naturally connects all parts of the numerical simulation process and gives rise to novel *fully adaptive algorithms*. We also theoretically address the question of convergence of the new fully adaptive algorithms. We identify theoretical conditions so that the error diminishes at each adaptive loop iteration by a contraction factor and we in particular derive a guaranteed error reduction factor in model cases. We have also proved a numerical optimality of the derived algorithms in model cases in the sense that, up to a generic constant, the smallest possible computational effort to achieve the given accuracy is needed.

3.5. Safe and correct programming

Finally, we concentrate on the issue of computer implementation of scientific computing programs. Increasing complexity of algorithms for modern scientific computing makes it a major challenge to implement them in the traditional imperative languages popular in the community. As an alternative, the computer science community provides theoretically sound tools for *safe* and *correct programming*. We explore here the use of these tools to design generic solutions for the implementation of the class of scientific computing software that we deal with. Our focus ranges from high-level programming via *functional programming* with OCAML through safe and easy parallelism via *skeleton parallel programming* with SKLML to proofs of correctness of numerical algorithms and programs via *mechanical proofs* with COQ.

SERPICO Project-Team

3. Research Program

3.1. Statistics and algorithms for computational microscopy

Fluorescence microscopy limitations are due to the optical aberrations, the resolution of the microscopy system, and the photon budget available for the biological specimen. Hence, new concepts have been defined to address challenging image restoration and molecule detection problems while preserving the integrity of samples. Accordingly, the main stream regarding denoising, deconvolution, registration and detection algorithms advocates appropriate signal processing framework to improve spatial resolution, while at the same time pushing the illumination to extreme low levels in order to limit photo-damages and phototoxicity. As a consequence, the question of adapting cutting-edge signal denoising and deconvolution, object detection, and image registration methods to 3D fluorescence microscopy imaging has retained the attention of several teams over the world.

In this area, the SERPICO team has developed a strong expertise in key topics in computational imaging including image denoising and deconvolution, object detection and multimodal image registration. Several algorithms proposed by the team outperformed the state-of-the-art results, and some developments are compatible with "high-throughput microscopy" and the processing of several hundreds of cells. We especially promoted non local, non-parametric and patch-based methods to solve well-known inverse problems or more original reconstruction problems. A recent research direction consists in adapting the deep learning concept to solve challenging detection and reconstruction problems in microscopy. We have investigated convolution neural networks to detect small macromolecules in 3D noisy electron images with promising results. The next step consists in proposing smart paradigms and architectures to save memory and computations.

More generally, many inverse problems and image processing become intractable with modern 3D microscopy, because very large temporal series of volumes (200 to 1000 images per second for one 3D stack) are acquired for several hours. Novel strategies are needed for 3D image denoising, deconvolution and reconstruction since computation is extremely heavy. Accordingly, we will adapt the estimator aggregation approach developed for optical flow computation to meet the requirements of 3D image processing. We plan to investigate regularization-based aggregation energy over super-voxels to reduce complexity, combined to modern optimization algorithms. Finally, we will design parallelized algorithms that fast process 3D images, perform energy minimization in few seconds per image, and run on low-cost graphics processor boards (GPU).

3.2. From image data to motion descriptors: trajectory computation and dynamics analysis

Several particle tracking methods for intracellular analysis have been tailored to cope with different types of cellular and subcellular motion down to Brownian single molecule behavior. Many algorithms were carefully evaluated on the particle tracking challenge dataset published in the Nature Methods journal in 2014. Actually, there is no definitive solution to the particle tracking problem which remains application-dependent in most cases. The work of SERPICO in particle motion analysis is significant in multiple ways, and inserts within a very active international context. One of the remaining key open issues is the tracking of objects with heterogeneous movements in crowded configurations. Moreover, particle tracking methods are not always adapted for motion analysis, especially when the density of moving features hampers the individual extraction of objects of interest undergoing complex motion. Estimating flow fields can be more appropriate to capture the complex dynamics observed in biological sequences. The existing optical flow methods can be classified into two main categories: i/ local methods impose a parametric motion model (e.g. local translation) in a given neighborhood; ii/ global methods estimate the dense motion field by minimizing a global energy functional composed of a data term and a regularization term.

The SERPICO team has developed a strong expertise in key topics, especially in object tracking for fluorescence microscopy, optical flow computation and high-level analysis of motion descriptors and trajectories. Several algorithms proposed by the team are very competitive when compared to the state-of-the-art results, and our new paradigms offer promising ways for molecule traffic quantification and analysis. Amongst the problems that we currently address, we can mention: computation of 3D optical flow for large-size images, combination of two frame-based differential methods and sparse sets of trajectories, detection and analysis of unexpected local motion patterns in global coherent collective motion. Development of efficient numerical schemes will be central in the future but visualization methods are also crucial for evaluation and quality assessment. Another direction of research consists in exploiting deep learning to 3D optical flow so as to develop efficient numerical schemes that naturally capture complex motion patterns. Investigation in machine learning and statistics will be actually conducted in the team in the two first research axes to address a large range of inverse problems in bioimaging. Deep learning is an appealing approach since expertise of biologists, via iterative annotation of training data, will be included in the design of image analysis schemes.

3.3. Biological and biophysical models and spatial statistics for quantitative bioimaging

A number of stochastic mathematical models were proposed to describe various intracellular trafficking, where molecules and proteins are transported to their destinations via free diffusion, subdiffusion and ballistic motion representing movements along the cytoskeleton networks assisted by molecular motors. Accordingly, the study of diffusion and stochastic dynamics has known a growing interest in bio-mathematics, biophysics and cell biology with the popularization of fluorescence dynamical microscopy and super-resolution imaging. In this area, the competing teams mainly studied MSD and fluorescence correlation spectroscopy methods.

In the recent period, the SERPICO team achieved important results for diffusion-related dynamics involved in exocytosis mechanisms. Robustness to noise has been well investigated, but robustness to environmental effects has yet to be effectively achieved. Particular attention has been given to the estimation of particle motion regime changes, but the available results are still limited for analyzing short tracks. The analysis of spatiotemporal molecular interactions from set of 3D computed trajectories or motion vector fields (e.g., co-alignment) must be investigated to fully quantify specific molecular machineries. We have already made efforts in that directions this year (e.g., for colocalization) but important experiments are required to make our preliminary algorithms reliable enough and well adapted to specific transport mechanisms.

Accordingly, we will study quantification methods to represent interactions between molecules and trafficking around three lines of research. First, we will focus on 3D space-time global and local object-based co-orientation and co-alignment methods, in the line of previous work on colocalization, to quantify interactions between molecular species. In addition, given N tracks associated to N molecular species, interaction descriptors, dynamics models and stochastic graphical models representing molecular machines will be studied in the statistical data assimilation framework. Second, we will analyse approaches to estimate molecular mobility, active transport and motion regime changes from computed trajectories in the Lagrangian and Eulerian settings. We will focus on the concept of super-resolution to provide spatially high-resolved maps of diffusion and active transport parameters based on stochastic biophysical models and sparse image representation. Third, we plan to extend the aggregation framework dedicated to optical flow to the problem of diffusion-transport estimation. Finally, we will investigate data assimilation methods to better combine algorithms, models, and experiments in an iterative and virtuous circle. The overview of ultrastructural organization will be achieved by additional 3D electron microscopy technologies.

SISTM Project-Team

3. Research Program

3.1. Mechanistic learning

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarizing the dynamics over time in term of slopes of the trajectories [56]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analyzing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [48]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [47].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [40]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [39].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [51], [59], [43], [49], [44], [58], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

Furthermore, the availability of data for each individual (see below) leads to a new challenge in this area. The structural model can easily be much more complex and the observation model may need to integrate much more markers.

3.2. High-dimensional statistical learning

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have became increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have became new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network [0]. As an example, in a phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production,

[0] see online at http://www.eurocoord.net

gene expression and viral genome change led to a 20 Go database [55]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these data-sets.

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high-dimensional data, the methodological challenge arises from the fact that data-sets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [52]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [57]. The integrative analysis of large data-sets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [45]. The comprehensive analysis of these large data-sets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

STEEP Project-Team

3. Research Program

3.1. Development of numerical systemic models (economy / society / environment) at local scales

The problem we consider is intrinsically interdisciplinary: it draws on social sciences, ecology or science of the planet. The modeling of the considered phenomena must take into account many factors of different nature which interact with varied functional relationships. These heterogeneous dynamics are *a priori* nonlinear and complex: they may have saturation mechanisms, threshold effects, and may be density dependent. The difficulties are compounded by the strong interconnections of the system (presence of important feedback loops) and multi-scale spatial interactions. Environmental and social phenomena are indeed constrained by the geometry of the area in which they occur. Climate and urbanization are typical examples. These spatial processes involve proximity relationships and neighborhoods, like for example, between two adjacent parcels of land, or between several macroscopic levels of a social organization. The multi-scale issues are due to the simultaneous consideration in the modeling of actors of different types and that operate at specific scales (spatial and temporal). For example, to properly address biodiversity issues, the scale at which we must consider the evolution of rurality is probably very different from the one at which we model the biological phenomena.

In this context, to develop flexible integrated systemic models (upgradable, modular, ...) which are efficient, realistic and easy to use (for developers, modelers and end users) is a challenge in itself. What mathematical representations and what computational tools to use? Nowadays many tools are used: for example, cellular automata (e.g. in the LEAM model), agent models (e.g. URBANSIM [0]), system dynamics (e.g. World3), large systems of ordinary equations (e.g. equilibrium models such as TRANUS), and so on. Each of these tools has strengths and weaknesses. Is it necessary to invent other representations? What is the relevant level of modularity? How to get very modular models while keeping them very coherent and easy to calibrate? Is it preferable to use the same modeling tools for the whole system, or can we freely change the representation for each considered subsystem? How to easily and effectively manage different scales? (difficulty appearing in particular during the calibration process). How to get models which automatically adapt to the granularity of the data and which are always numerically stable? (this has also a direct link with the calibration processes and the propagation of uncertainties). How to develop models that can be calibrated with reasonable efforts, consistent with the (human and material) resources of the agencies and consulting firms that use them?

Before describing our research axes, we provide a brief overview of the types of models that we are or will be working with. As for LUTI (Land Use and Transportation Integrated) modeling, we have been using the TRANUS model since the start of our group. It is the most widely used LUTI model, has been developed since 1982 by the company Modelistica, and is distributed *via* Open Source software. TRANUS proceeds by solving a system of deterministic nonlinear equations and inequalities containing a number of economic parameters (e.g. demand elasticity parameters, location dispersion parameters, etc.). The solution of such a system represents an economic equilibrium between supply and demand.

On the other hand, the scientific domains related to ecosystem services and ecological accounting are much less mature than the one of urban economy from a modelling point of view (as a consequence of our more limited knowledge of the relevant complex processes and/or more limited available data). Nowadays, the community working on ecological accounting develops statistical models based on the enforcement of the mass conservation constraint for accounting for material fluxes through a territorial unit or a supply chain, relying on more or less simple data correlations when the relevant data is missing; the overall modelling makes heavy use of more or less sophisticated linear algebra and constrained optimization techniques. The

[0] http://www.urbansim.org

ecosystem service community has been using statical models too, but is also developing more sophisticated models based for example on system dynamics, multi-agent type simulations or cellular models. In the ESNET project, STEEP has worked in particular on a land use/land cover change (LUCC) modelling environments (Dinamica [0]) which belongs to the category of spatially explicit statistical models.

In the following, our two main research axes are described, from the point of view of applied mathematical development. The domains of application of this research effort is described in the application section, where some details about the context of each field is given.

3.2. Model calibration and validation

The overall calibration of the parameters that drive the equations implemented in the above models is a vital step. Theoretically, as the implemented equations describe e.g. socio-economic phenomena, some of these parameters should in principle be accurately estimated from past data using econometrics and statistical methods like regressions or maximum likelihood estimates, e.g. for the parameters of logit models describing the residential choices of households. However, this theoretical consideration is often not efficient in practice for at least two main reasons. First, the above models consist of several interacting modules. Currently, these modules are typically calibrated independently; this is clearly sub-optimal as results will differ from those obtained after a global calibration of the interaction system, which is the actual final objective of a calibration procedure. Second, the lack of data is an inherent problem.

As a consequence, models are usually calibrated by hand. The calibration can typically take up to 6 months for a medium size LUTI model (about 100 geographic zones, about 10 sectors including economic sectors, population and employment categories). This clearly emphasizes the need to further investigate and at least semi-automate the calibration process. Yet, in all domains STEEP considers, very few studies have addressed this central issue, not to mention calibration under uncertainty which has largely been ignored (with the exception of a few uncertainty propagation analyses reported in the literature).

Besides uncertainty analysis, another main aspect of calibration is numerical optimization. The general state-of-the-art on optimization procedures is extremely large and mature, covering many different types of optimization problems, in terms of size (number of parameters and data) and type of cost function(s) and constraints. Depending on the characteristics of the considered models in terms of dimension, data availability and quality, deterministic or stochastic methods will be implemented. For the former, due to the presence of non-differentiability, it is likely, depending on their severity, that derivative free control methods will have to be preferred. For the latter, particle-based filtering techniques and/or metamodel-based optimization techniques (also called response surfaces or surrogate models) are good candidates.

These methods will be validated, by performing a series of tests to verify that the optimization algorithms are efficient in the sense that 1) they converge after an acceptable computing time, 2) they are robust and 3) that the algorithms do what they are actually meant to. For the latter, the procedure for this algorithmic validation phase will be to measure the quality of the results obtained after the calibration, i.e. we have to analyze if the calibrated model fits sufficiently well the data according to predetermined criteria.

To summarize, the overall goal of this research axis is to address two major issues related to calibration and validation of models: (a) defining a calibration methodology and developing relevant and efficient algorithms to facilitate the parameter estimation of considered models; (b) defining a validation methodology and developing the related algorithms (this is complemented by sensitivity analysis, see the following section). In both cases, analyzing the uncertainty that may arise either from the data or the underlying equations, and quantifying how these uncertainties propagate in the model, are of major importance. We will work on all those issues for the models of all the applied domains covered by STEEP.

3.3. Sensitivity analysis

[0] http://www.csr.ufmg.br/dinamica/

A sensitivity analysis (SA) consists, in a nutshell, in studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs. It is complementary to an uncertainty analysis, which focuses on quantifying uncertainty in model output. SA's can be useful for several purposes, such as guiding model development and identifying the most influential model parameters and critical data items. Identifying influential model parameters may help in divising metamodels (or, surrogate models) that approximate an original model and may be simulated, calibrated, or analyzed more efficiently. As for detecting critical data items, this may indicate for which type of data more effort must be spent in the data collection process in order to eventually improve the model's reliability. Finally, SA can be used as one means for validating models, together with validation based on historical data (or, put simply, using training and test data) and validation of model parameters and outputs by experts in the respective application area.

The first two applications of SA are linked to model calibration, discussed in the previous section. Indeed, prior to the development of the calibration tools, one important step is to select the significant or sensitive parameters and to evaluate the robustness of the calibration results with respect to data noise (stability studies). This may be performed through a global sensitivity analysis, e.g. by computation of Sobol's indices. Many problems had been to be circumvented e.g. difficulties arising from dependencies of input variables, variables that obey a spatial organization, or switch inputs. We take up on current work in the statistics community on SA for these difficult cases.

As for the third application of SA, model validation, a preliminary task bears on the propagation of uncertainties. Identifying the sources of uncertainties and their nature is crucial to propagate them via Monte Carlo techniques. To make a Monte Carlo approach computationally feasible, it is necessary to develop specific metamodels. Both the identification of the uncertainties and their propagation require a detailed knowledge of the data collection process; these are mandatory steps before a validation procedure based on SA can be implemented. First, we focus on validating LUTI models, starting with the CITiES ANR project: here, an SA consists in defining various land use policies and transportation scenarios and in using these scenarios to test the integrated land use and transportation model. Current approaches for validation by SA consider several scenarios and propose various indicators to measure the simulated changes. We work towards using sensitivity indices based on functional analysis of variance, which allow us to compare the influence of various inputs on the indicators. For example it allow the comparison of the influences of transportation and land use policies on several indicators.

3.4. Global systemic risks

Modern societies are characterized by a very high level of global interconnections between many sectors of economic, social and political activity, as well as by the environmental impacts produced by these activities and their consequences for human societies themselves. The resulting generalized interdependences induce intrinsic, or "systemic", risks of instability. These risks constitute serious threats for the global socio-ecological system, and the issue of a potential global collapse is part of the threats to be analyzed.

Global systemic risks directly relate to the STEEP team project, i.e., the question of sustainability at various spatial scales. The ability of socio-ecosystems, local communities, nation States and the international community to address these risks is a key factor in a sustainability perspective. However, the topic of global systemic risks was not until recently within the scope of the research activity of the team.

Within academia, the activity of several research institutes is devoted to these risks, with a strong contribution of social sciences. The most representative are probably the *Global Systemic Risk Institute* of Princeton [0], the *Center for Risk Studies* of Cambridge (UK) Cambridge [0], and the *Risk Center* of Zürich [0]. Various teams are also active on these themes, but with a more sectorial focus.

[0] https://risk.princeton.edu/
[0] https://www.jbs.cam.ac.uk/faculty-research/centres/risk/
[0] https://riskcenter.ethz.ch/

From the point of view of the main processes at work, global systemic risks can be grouped into two categories

1. Risks related to long term mean trends (several decades). For the most part, they are generated by the growing tension between our use of resources and pollution production and our (semi)-natural environment capacity to absorb the associated impacts. They also emerge from the back-reaction of these environmental changes on socio-ecosystems. These risks are underlined and amplified by specific historical, socio-politic and economic dynamics.
2. Risks related to systemic contagion effect, on much shorter terms (weeks or months), but sporadic and random. This type of risk is directly driven by the high level of interconnection of various large scale human activities, to intrinsic instabilities whose very existence is directly tied to these interconnections, and to the potential propagation of these instabilities in all sectors of activity, in a domino-like effect. These risks are amplified by current geopolical dynamics and by the deepening of the various environmental crises.

In such a context, and due to its areas of expertise, the STEEP team has a nearly unique ability and opportunity to make a significant contribution to these questions, most particularly on the modelling front.

The World3 model is without any doubt the most representative of the first category of risks. It was developed by Meadows and coworkers for their famous report on the limits to growth [18], [20]. The reinvestigation by [22], [23] and re-discussion by [13] have renewed the interest for this model, while raising more pinpointed questions on the robustness and validity of the conclusions drawn from it. We have started to address these questions on three different fronts:

1. Through an analysis of the parameterization choices performed in the mode. In practice the team has undertaken a sensitivity analysis that is substantially more comprehensive than previous attempts in this direction.
2. Through an analysis of the modelling choices performed by the Meadows group. This will consist in a partial sectoral and spatial disaggregation of the model.
3. Through some elements of epistemological analysis.

As a matter of fact, two internships have already been devoted to the sensitivity analysis of the model. This is now pursued through a PhD thesis, started last Fall. The student (Mathilde Duplessix) will also address the other two points in the course of her PhD.

The main practical interest of this work is related to the possibility of distinguishing a potential onset of collapse in the first or second half ot the century. These two broad options imply different mitigation/adaptation strategies, that need to be correctly anticipated.

On the front of systemic contagion risks, and although a comprehensive analysis of the whole range of potential risks is impossible in an exploratory phase, the nexus energy/finance/supply chains plays a particular role in our societies and present a specific level of criticity. Some sectorial (and even cross-sectorial) aspects of this nexust have been discussed in the literature (e.g., [16], [17], [14]), but apparently no global, generic has been produced so far. Such a model would constitute by itself a remarkable breakthrough in this topic.

Funding for these research objectives has been obtained this year, in the form of an Inria “Action Exploratoire” project which officially starts in January 2020. This funding covers a PhD (Louis Delannoy, starting in January 2020) and a post-doctoral position (scheduled to start in 2021).

TONUS Project-Team

3. Research Program

3.1. Kinetic models for plasmas

The fundamental model for plasma physics is the coupled Vlasov-Maxwell kinetic model: the Vlasov equation describes the distribution function of particles (ions and electrons), while the Maxwell equations describe the electromagnetic field. In some applications, it may be necessary to take relativistic particles into account, which leads to consider the relativistic Vlasov equation, even if in general, tokamak plasmas are supposed to be non-relativistic. The distribution function of particles depends on seven variables (three for space, three for the velocity and one for time), which yields a huge amount of computation. To these equations we must add several types of source terms and boundary conditions for representing the walls of the tokamak, the applied electromagnetic field that confines the plasma, fuel injection, collision effects, etc.

Tokamak plasmas possess particular features, which require developing specialized theoretical and numerical tools.

Because the magnetic field is strong, the particle trajectories have a very fast rotation around the magnetic field lines. A full resolution would require a prohibitive amount of computation. It is necessary to develop reduced models for large magnetic fields in order to obtain tractable calculations. The resulting model is called a gyrokinetic model. It allows us to reduce the dimensionality of the problem. Such models are implemented in GYSELA and Selalib.

On the boundary of the plasma, the collisions can no more be neglected. Fluid models, such as MagnetoHydroDynamics (MHD) become again relevant. For the good operation of the tokamak, it is necessary to control MHD instabilities that arise at the plasma boundary. Computing these instabilities requires special implicit numerical discretizations with excellent long time behavior.

In addition to theoretical modelling tools, it is necessary to develop numerical schemes adapted to kinetic, gyrokinetic and fluid models. Three kinds of methods are studied in TONUS: Particle-In-Cell (PIC) methods, semi-Lagrangian and fully Eulerian approaches.

3.1.1. Gyrokinetic models: theory and approximation

In most phenomena where oscillations are present, we can establish a three-model hierarchy: (i) the model parameterized by the oscillation period, (ii) the limit model and (iii) the two-scale model, possibly with its corrector. In a context where one wishes to simulate such a phenomenon where the oscillation period is small and the oscillation amplitude is not small, it is important to have numerical methods based on an approximation of the two-scale model. If the oscillation period varies significantly over the domain of simulation, it is important to have numerical methods that approximate properly and effectively the model parameterized by the oscillation period and the two-scale model. Implementing two-scale numerical methods (for instance by Frénod et al. [27]) is based on a numerical approximation of the Two-Scale model. These are called of order 0. A Two-Scale Numerical Method is called of order 1 if it incorporates information from the corrector and from the equation of which this corrector is a solution. If the oscillation period varies between very small values and values of order 1, it is necessary to have new types of numerical schemes (Two-Scale Asymptotic Preserving Schemes of order 1 or TSAPS) that preserve the asymptotics between the model parameterized by the oscillation period and the Two-Scale model with its corrector. A first work in this direction has been initiated by Crouseilles et al. [26].

3.1.2. Semi-Lagrangian schemes

The Strasbourg team has a long and recognized experience in numerical methods for Vlasov-type equations. We are specialized in both particle and phase space solvers for the Vlasov equation: Particle-in-Cell (PIC) methods and semi-Lagrangian methods. We also have a long-standing collaboration with CEA Cadarache for the development of the GYSELA software for gyrokinetic tokamak plasmas.

The Vlasov and the gyrokinetic models are partial differential equations that express the transport of the distribution function in the phase space. In the original Vlasov case, the phase space is the six-dimension position-velocity space. For the gyrokinetic model, the phase space is five-dimensional because we consider only the parallel velocity in the direction of the magnetic field and the gyrokinetic angular velocity instead of three velocity components.

A few years ago, Eric Sonnendrücker and his collaborators introduced a new family of methods for solving transport equations in the phase space. This family of methods are the semi-Lagrangian methods. The principle of these methods is to solve the equation on a grid of the phase space. The grid points are transported with the flow of the transport equation for a time step and interpolated back periodically onto the initial grid. The method is then a mix of particle Lagrangian methods and Eulerian methods. The characteristics can be solved forward or backward in time leading to the Forward Semi-Lagrangian (FSL) or Backward Semi-Lagrangian (BSL) schemes. Conservative schemes based on this idea can be developed and are called Conservative Semi-Lagrangian (CSL).

GYSELA is a 5D full gyrokinetic code based on a classical backward semi-Lagrangian scheme (BSL) [31] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [28].

More recently, we have started to apply the semi-Lagrangian methods to more general kinetic equations. Indeed, most of the conservation laws of physics can be represented by a kinetic model with a small set of velocities and relaxation source terms [4]. Compressible fluids or MHD equations have such representations. Semi-Lagrangian methods then become a very appealing and efficient approach for solving these equations.

3.1.3. PIC methods

Historically PIC methods have been very popular for solving the Vlasov equations. They allow solving the equations in the phase space at a relatively low cost. The main disadvantage of this approach is that, due to its random aspect, it produces an important numerical noise that has to be controlled in some way, for instance by regularizations of the particles, or by divergence correction techniques in the Maxwell solver. We have a long-standing experience in PIC methods and we started implementing them in Selalib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [25].

3.2. Fluid and reduced kinetic models for plasmas

As already said, kinetic plasmas computer simulations are very intensive, because of the gyrokinetic turbulence. In some situations, it is possible to make assumptions on the shape of the distribution function that simplify the model. We obtain in this way a family of fluid or reduced models.

Assuming that the distribution function has a Maxwellian shape, for instance, we obtain the MagnetoHydroDynamic (MHD) model. It is physically valid only in some parts of the tokamak (at the edges for instance). The fluid model is generally obtained from the hypothesis that the collisions between particles are strong.

But the reduction is not necessarily a consequence of collisional effects. Indeed, even without collisions, the plasma may still relax to an equilibrium state over sufficiently long time scales (Landau damping effect).

In the fluid or reduced-kinetic regions, the approximation of the distribution function could require fewer data while still achieving a good representation, even in the collisionless regime.

Therefore, a fluid or a reduced model is a model where the explicit dependency on the velocity variable is removed. In a more mathematical way, we consider that in some regions of the plasma, it is possible to exhibit a (preferably small) set of parameters α that allows us to describe the main properties of the plasma with a generalized "Maxwellian" M. Then

$$f(x, v, t) = M(\alpha(x, t), v).$$

In this case it is sufficient to solve for $\alpha(x,t)$. Generally, the vector α is the solution of a first order hyperbolic system.

Another way to reduce the model is to try to find an abstract kinetic representation with an as small as possible set of kinetic velocities. The kinetic approach has then only a mathematical meaning. It allows solving very efficiently many equations of physics.

3.2.1. Numerical schemes

As previously indicated, an efficient method for solving the reduced models is the Discontinuous Galerkin (DG) approach. It is possible to make it of arbitrary order. It requires limiters when it is applied to nonlinear PDEs occurring for instance in fluid mechanics. But the reduced models that we intend to write are essentially linear. The nonlinearity is concentrated in a few coupling source terms.

In addition, this method, when written in a special set of variables, called the entropy variables, has nice properties concerning the entropy dissipation of the model. It opens the door to constructing numerical schemes with good conservation properties and no entropy dissipation, as already used for other systems of PDEs [32], [24], [30], [29].

3.2.2. Matrix-free implicit schemes

In tokamaks, the reduced model generally involves many time scales. Among these time scales, many of them, associated to the fastest waves, are not relevant. In order to filter them out, it is necessary to adopt implicit solvers in time. When the reduced model is based on a kinetic interpretation, it is possible to construct implicit schemes that do not impose solving costly linear systems. In addition the resulting solver is stable even at a very high CFL (Courant Friedrichs Lax) number.

3.3. Electromagnetic solvers

Precise resolution of the electromagnetic fields is essential for proper plasma simulation. Thus it is important to use efficient solvers for the Maxwell systems and its asymptotics: Poisson equation and magnetostatics.

The proper coupling of the electromagnetic solver with the Vlasov solver is also crucial for ensuring conservation properties and stability of the simulation.

Finally, plasma physics implies very different time scales. It is thus very important to develop implicit Maxwell solvers and Asymptotic Preserving (AP) schemes in order to obtain good behavior on long time scales.

3.3.1. Coupling

The coupling of the Maxwell equations to the Vlasov solver requires some precautions. The most important one is to control the charge conservation errors, which are related to the divergence conditions on the electric and magnetic fields. We will generally use divergence correction tools for hyperbolic systems presented for instance in [23] (and the references therein).

3.3.2. Implicit solvers

As already pointed out, in a tokamak, the plasma presents several different space and time scales. It is not possible in practice to solve the initial Vlasov-Maxwell model. It is first necessary to establish asymptotic models by letting some parameters (such as the Larmor frequency or the speed of light) tend to infinity. This is the case for the electromagnetic solver and this requires implementing implicit time solvers in order to efficiently capture the stationary state, the solution of the magnetic induction equation or the Poisson equation.

XPOP Project-Team

3. Research Program

3.1. Scientific positioning

"Interfaces" is the defining characteristic of XPOP:

The interface between statistics, probability and numerical methods. Mathematical modelling of complex biological phenomena require to combine numerical, stochastic and statistical approaches. The CMAP is therefore the right place to be for positioning the team at the interface between several mathematical disciplines.

The interface between mathematics and the life sciences. The goal of XPOP is to bring the right answers to the right questions. These answers are mathematical tools (statistics, numerical methods, etc.), whereas the questions come from the life sciences (pharmacology, medicine, biology, etc.). This is why the point of XPOP is not to take part in mathematical projects only, but also pluridisciplinary ones.

The interface between mathematics and software development. The development of new methods is the main activity of XPOP. However, new methods are only useful if they end up being implemented in a software tool. On one hand, a strong partnership with Lixoft (the spin-off company who continue developing MONOLIX) allows us to maintaining this positioning. On the other hand, several members of the team are very active in the R community and develop widely used packages.

3.2. The mixed-effects models

Mixed-effects models are statistical models with both fixed effects and random effects. They are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p_Y(y_i;\psi_i)$ be the probability distribution of y_i, where ψ_i is a vector of parameters.

In a population framework, the vector of parameters ψ_i is assumed to be drawn from a population distribution $p_\Psi(\psi_i;\theta)$ where θ is a vector of population parameters.

Then, the probabilistic model is the joint probability distribution

$$p(y_i,\psi_i;\theta) = p_Y(y_i|\psi_i)p_\Psi(\psi_i;\theta) \tag{65}$$

To define a model thus consists in defining precisely these two terms.

In most applications, the observed data y_i are continuous longitudinal data. We then assume the following representation for y_i:

$$y_{ij} = f(t_{ij},\psi_i) + g(t_{ij},\psi_i)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i. \tag{66}$$

Here, y_{ij} is the observation obtained from subject i at time t_{ij}. The residual errors (ε_{ij}) are assumed to be standardized random variables (mean zero and variance 1). The residual error model is represented by function g in model (2).

Function f is usually the solution to a system of ordinary differential equations (pharmacokinetic/pharmacodynamic models, etc.) or a system of partial differential equations (tumor growth, respiratory system, etc.). This component is a fundamental component of the model since it defines the prediction of the observed kinetics for a given set of parameters.

The vector of individual parameters ψ_i is usually function of a vector of population parameters $\psi_{\rm pop}$, a vector of random effects $\eta_i \sim \mathcal{N}(0, \Omega)$, a vector of individual covariates c_i (weight, age, gender, ...) and some fixed effects β.

The joint model of y and ψ depends then on a vector of parameters $\theta = (\psi_{\rm pop}, \beta, \Omega)$.

3.3. Computational Statistical Methods

Central to modern statistics is the use of probabilistic models. To relate these models to data requires the ability to calculate the probability of the observed data: the likelihood function, which is central to most statistical methods and provides a principled framework to handle uncertainty.

The emergence of computational statistics as a collection of powerful and general methodologies for carrying out likelihood-based inference made complex models with non-standard data accessible to likelihood, including hierarchical models, models with intricate latent structure, and missing data.

In particular, algorithms previously developed by POPIX for mixed effects models, and today implemented in several software tools (especially MONOLIX) are part of these methods:

- the adaptive Metropolis-Hastings algorithm allows one to sample from the conditional distribution of the individual parameters $p(\psi_i|y_i; c_i, \theta)$,
- the SAEM algorithm is used to maximize the observed likelihood $\mathcal{L}(\theta; y) = p(y; \theta)$,
- Importance Sampling Monte Carlo simulations provide an accurate estimation of the observed log-likelihood $\log(\mathcal{L}(\theta; y))$.

Computational statistics is an area which remains extremely active today. Recently, one can notice that the incentive for further improvements and innovation comes mainly from three broad directions: the high dimensional challenge, the quest for adaptive procedures that can eliminate the cumbersome process of tuning "by hand" the settings of the algorithms and the need for flexible theoretical support, arguably required by all recent developments as well as many of the traditional MCMC algorithms that are widely used in practice.

Working in these three directions is a clear objective for XPOP.

3.4. Markov Chain Monte Carlo algorithms

While these Monte Carlo algorithms have turned into standard tools over the past decade, they still face difficulties in handling less regular problems such as those involved in deriving inference for high-dimensional models. One of the main problems encountered when using MCMC in this challenging settings is that it is difficult to design a Markov chain that efficiently samples the state space of interest.

The Metropolis-adjusted Langevin algorithm (MALA) is a Markov chain Monte Carlo (MCMC) method for obtaining random samples from a probability distribution for which direct sampling is difficult. As the name suggests, MALA uses a combination of two mechanisms to generate the states of a random walk that has the target probability distribution as an invariant measure:

1. new states are proposed using Langevin dynamics, which use evaluations of the gradient of the target probability density function;
2. these proposals are accepted or rejected using the Metropolis-Hastings algorithm, which uses evaluations of the target probability density (but not its gradient).

Informally, the Langevin dynamics drives the random walk towards regions of high probability in the manner of a gradient flow, while the Metropolis-Hastings accept/reject mechanism improves the mixing and convergence properties of this random walk.

Several extensions of MALA have been proposed recently by several authors, including fMALA (fast MALA), AMALA (anisotropic MALA), MMALA (manifold MALA), position-dependent MALA (PMALA), ...

MALA and these extensions have demonstrated to represent very efficient alternative for sampling from high dimensional distributions. We therefore need to adapt these methods to general mixed effects models.

3.5. Parameter estimation

The Stochastic Approximation Expectation Maximization (SAEM) algorithm has shown to be extremely efficient for maximum likelihood estimation in incomplete data models, and particularly in mixed effects models for estimating the population parameters. However, there are several practical situations for which extensions of SAEM are still needed:

High dimensional model: a complex physiological model may have a large number of parameters (in the order of 100). Then several problems arise:

- when most of these parameters are associated with random effects, the MCMC algorithm should be able to sample, for each of the N individuals, parameters from a high dimensional distribution. Efficient MCMC methods for high dimensions are then required.
- Practical identifiability of the model is not ensured with a limited amount of data. In other words, we cannot expect to be able to properly estimate all the parameters of the model, including the fixed effects and the variance-covariance matrix of the random effects. Then, some random effects should be removed, assuming that some parameters do not vary in the population. It may also be necessary to fix the value of some parameters (using values from the literature for instance). The strategy to decide which parameters should be fixed and which random effects should be removed remains totally empirical. XPOP aims to develop a procedure that will help the modeller to take such decisions.

Large number of covariates: the covariate model aims to explain part of the inter-patient variability of some parameters. Classical methods for covariate model building are based on comparisons with respect to some criteria, usually derived from the likelihood (AIC, BIC), or some statistical test (Wald test, LRT, etc.). In other words, the modelling procedure requires two steps: first, all possible models are fitted using some estimation procedure (e.g. the SAEM algorithm) and the likelihood of each model is computed using a numerical integration procedure (e.g. Monte Carlo Importance Sampling); then, a model selection procedure chooses the "best" covariate model. Such a strategy is only possible with a reduced number of covariates, i.e., with a "small" number of models to fit and compare.

As an alternative, we are thinking about a Bayesian approach which consists of estimating simultaneously the covariate model and the parameters of the model in a single run. An (informative or uninformative) prior is defined for each model by defining a prior probability for each covariate to be included in the model. In other words, we extend the probabilistic model by introducing binary variables that indicate the presence or absence of each covariate in the model. Then, the model selection procedure consists of estimating and maximizing the conditional distribution of this sequence of binary variables. Furthermore, a probability can be associated to any of the possible covariate models.

This conditional distribution can be estimated using an MCMC procedure combined with the SAEM algorithm for estimating the population parameters of the model. In practice, such an approach can only deal with a limited number of covariates since the dimension of the probability space to explore increases exponentially with the number of covariates. Consequently, we would like to have methods able to find a small number of variables (from a large starting set) that influence certain parameters in populations of individuals. That means that, instead of estimating the conditional distribution of all the covariate models as described above, the algorithm should focus on the most likely ones.

Fixed parameters: it is quite frequent that some individual parameters of the model have no random component and are purely fixed effects. Then, the model may not belong to the exponential family anymore and the original version of SAEM cannot be used as it is. Several extensions exist:

- introduce random effects with decreasing variances for these parameters,
- introduce a prior distribution for these fixed effects,
- apply the stochastic approximation directly on the sequence of estimated parameters, instead of the sufficient statistics of the model.

None of these methods always work correctly. Furthermore, what are the pros and cons of these methods is not clear at all. Then, developing a robust methodology for such model is necessary.

Convergence toward the global maximum of the likelihood: convergence of SAEM can strongly depend on thie initial guess when the observed likelihood has several local maxima. A kind of simulated annealing version of SAEM was previously developed and implemented in MONOLIX. The method works quite well in most situations but there is no theoretical justification and choosing the settings of this algorithm (i.e. how the temperature decreases during the iterations) remains empirical. A precise analysis of the algorithm could be very useful to better understand why it "works" in practice and how to optimize it.

Convergence diagnostic: Convergence of SAEM was theoretically demonstrated under very general hypothesis. Such result is important but of little interest in practice at the time to use SAEM in a finite amount of time, i.e. in a finite number of iterations. Some qualitative and quantitative criteria should be defined in order to both optimize the settings of the algorithm, detect a poor convergence of SAEM and evaluate the quality of the results in order to avoid using them unwisely.

3.6. Model building

Defining an optimal strategy for model building is far from easy because a model is the assembled product of numerous components that need to been evaluated and perhaps improved: the structural model, residual error model, covariate model, covariance model, etc.

How to proceed so as to obtain the best possible combination of these components? There is no magic recipe but an effort will be made to provide some qualitative and quantitative criteria in order to help the modeller for building his model.

The strategy to take will mainly depend on the time we can dedicate to building the model and the time required for running it. For relatively simple models for which parameter estimation is fast, it is possible to fit many models and compare them. This can also be done if we have powerful computing facilities available (e.g., a cluster) allowing large numbers of simultaneous runs.

However, if we are working on a standard laptop or desktop computer, model building is a sequential process in which a new model is tested at each step. If the model is complex and requires significant computation time (e.g., when involving systems of ODEs), we are constrained to limit the number of models we can test in a reasonable time period. In this context, it also becomes important to carefully choose the tasks to run at each step.

3.7. Model evaluation

Diagnostic tools are recognized as an essential method for model assessment in the process of model building. Indeed, the modeler needs to confront "his" model with the experimental data before concluding that this model is able to reproduce the data and before using it for any purpose, such as prediction or simulation for instance.

The objective of a diagnostic tool is twofold: first we want to check if the assumptions made on the model are valid or not ; then, if some assumptions are rejected, we want to get some guidance on how to improve the model.

As is the usual case in statistics, it is not because this "final" model has not been rejected that it is necessarily the "true" one. All that we can say is that the experimental data does not allow us to reject it. It is merely one of perhaps many models that cannot be rejected.

Model diagnostic tools are for the most part graphical, i.e., visual; we "see" when something is not right between a chosen model and the data it is hypothesized to describe. These diagnostic plots are usually based on the empirical Bayes estimates (EBEs) of the individual parameters and EBEs of the random effects: scatterplots of individual parameters versus covariates to detect some possible relationship, scatterplots of pairs of random effects to detect some possible correlation between random effects, plot of the empirical distribution of the random effects (boxplot, histogram,...) to check if they are normally distributed, ...

The use of EBEs for diagnostic plots and statistical tests is efficient with rich data, i.e. when a significant amount of information is available in the data for recovering accurately all the individual parameters. On the contrary, tests and plots can be misleading when the estimates of the individual parameters are greatly shrunk.

We propose to develop new approaches for diagnosing mixed effects models in a general context and derive formal and unbiased statistical tests for testing separately each feature of the model.

3.8. Missing data

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst.

Missing values occur for a variety of reasons: machines that fail, survey participants who do not answer certain questions, destroyed or lost data, dead animals, damaged plants, etc. Missing values are problematic since most statistical methods can not be applied directly on a incomplete data. Many progress have been made to properly handle missing values. However, there are still many challenges that need to be addressed in the future, that are crucial for the users.

- State of arts methods often consider the case of continuous or categorical data whereas real data are very often mixed. The idea is to develop a multiple imputation method based on a specific principal component analysis (PCA) for mixed data. Indeed, PCA has been used with success to predict (impute) the missing values. A very appealing property is the ability of the method to handle very large matrices with large amount of missing entries.
- The asymptotic regime underlying modern data is not any more to consider that the sample size increases but that both number of observations and number of variables are very large. In practice first experiments showed that the coverage properties of confidence areas based on the classical methods to estimate variance with missing values varied widely. The asymptotic method and the bootstrap do well in low-noise setting, but can fail when the noise level gets high or when the number of variables is much greater than the number of rows. On the other hand, the jackknife has good coverage properties for large noisy examples but requires a minimum number of variables to be stable enough.
- Inference with missing values is usually performed under the assumption of "Missing at Random" (MAR) values which means that the probability that a value is missing may depend on the observed data but does not depend on the missing value itself. In real data and in particular in data coming from clinical studies, both "Missing Non at Random" (MNAR) and MAR values occur. Taking into account in a proper way both types of missing values is extremely challenging but is worth investigating since the applications are extremely broad.

It is important to stress that missing data models are part of the general incomplete data models addressed by XPOP. Indeed, models with latent variables (i.e. non observed variables such as random effects in a mixed effects model), models with censored data (e.g. data below some limit of quantification) or models with dropout mechanism (e.g. when a subject in a clinical trial fails to continue in the study) can be seen as missing data models.

AGORA Project-Team

3. Research Program

3.1. Wireless network deployment

The deployment of networks has fostered a constant research effort for decades, continuously renewed by the evolution of networking technologies. Fundamentally, the deployment problem addresses the trade-off between the cost of the network to be minimized or fitted into a budget and the features and services provided by the system, that should reach a target level or be maximized. The variety of cost models and type of features gives rise to a wide scientific field. There are several cost factors of network infrastructure: components (number and capacity), energy, man power (installation and maintenance), etc. The features of the network matter as much as the metric to evaluate them. Coverage and capacity are basic features for wireless networks on which we will focus in the following. One recurrent question is therefore: What are the optimal number and position of network components to deploy so that a given territory is covered and enough networking capacity is provided?

Traditional telecommunication infrastructures were made of dedicated components, each of them providing a given set of functions. However, recently introduced paradigms yield issues on the deployment of network functions. Indeed, the last decade saw a trend towards adding more intelligence within the network. In the case of the access network, the concept of Cloud Radio Access Network (C-RAN) emerged. In the backhaul, the Evolved Packet Core (EPC) network can also benefit from virtualization techniques, as the convergence point for multiple access technologies, as imagined in the case of future 5G networks. The performance limits of a virtualized EPC remain unknown today: Is the delay introduced by this new architecture compatible with the requirements of the mobile applications? How to deploy the different network functions on generic hardware in order to maximize the quality of service?

Network component deployment. In this research direction, we address new issues of the optimal network deployment. In particular, we focus on the deployment of wireless sensor networks for environmental monitoring (e.g. atmospheric pollution). Most current air quality monitoring systems are using conventional measuring stations, equipped with multiple lab quality sensors. These systems are however massive, inflexible and expensive. An alternative – or complementary – solution is to use low-cost flexible wireless sensor networks. One of the main challenges is to introduce adequate models for the coverage of the phenomenon. Most of the state of the art considers a generic coverage formulation based on detection ranges which are not adapted to environmental sensing. For example, pollution propagation models should take into account the inherently stochastic weather conditions. An issue is to develop an adequate formulation and efficient integer linear programming (ILP) models and heuristics able to compute deployments at a relevant scale. In particular, it seems promising to adapt stochastic or robust optimization results of the operational research community in order to deals with uncertainty. Defining the quality of a coverage is also a modeling issue, which depends on the application considered. The detection of anomaly is close to a combinatorial problem. A more difficult objective is to deploy sensors in order to map the phenomenon by interpolation (or other reconstruction mechanisms). This challenge requires interdisciplinary research with fluid mechanics teams who develop numerical models of pollution propagation and practitioners like Atmo Auvergne-Rhône-Alpes.

Regarding the network connectivity, another challenge is to integrate suitable wireless link models accounting for the deployment environment. For example, modeling the integration of sensors in urban areas is challenging due to the presence of neighboring walls and obstacles, as well as moving vehicles and pedestrians that may induce field scattering. Also, the urban constraints and characteristics need to be carefully modeled and considered. Indeed, the urban environment yields constraints or facilities on the deployment of sensor nodes and gateways, such as their embedding within street furniture. Understanding the structure of these spatial constraints is necessary to develop efficient optimization methods able to compute on large scale scenarios.

Network function deployment. In this research direction, we do not address network virtualization per se, but the algorithmic and architectural challenges that virtualization brings in both radio access and core networks. As a first challenge, we focus on the evaluation of Cloud Radio Access Network solutions. The capacity of a C-RAN architecture and the way this compares to classical RAN is still an open question. The fact that C-RAN enables cooperation between the remote radio heads (RRH) served by the same base-band units (BBU) indicates an improved performance, but at the same time the resulting cells are much larger, which goes against the current trend of increasing capacity through the deployment of small cells. We propose to study the problem both from a user and a network perspective. On the user side, we use standard information theory tools, such as multiple-access channels to model C-RAN scenarios and understand their performance. On the network side, this translates in a resource allocation problem with cooperative base stations. We will extend our previous models for non-cooperative scenarios. Regarding the core network function deployment, we are interested in the specific case of Professional Mobile Radio (PMR) networks. These networks, used for public safety services and in scenarios like post-disaster relief, present the particularity of an EPC formed by a mobile wireless network. Due to its nature, the network can not be pre-planned, and the different EPC functions need to be autonomously deployed on the available network elements. We study the EPC function deployment problem as an optimization problem, constrained by the user capacity requests. User attachment mechanisms will also be proposed, adapted to the network function distribution, the global user demand, and the source/destination of the flows. These challenges are tackled as centralized optimization problems, then extended to the context of real-time decisions. Finally, in order to complete these theoretical works based on ILP models and heuristics, experiments using OpenAir Interface are used to evaluate our proposals.

3.2. Wireless data collection

With an anticipated 11-fold growth between 2014 and 2018, facing the growth of the mobile demand is the foremost challenge for mobile operators. In particular, a 100-fold increase in the number of supported connected devices, mostly newly connected objects with M2M trafic, is expected. A question therefore arises: how to cope with a dense set of M2M low bit rate traffics from energy and computing power constrained devices while classic cellular infrastructures are designed for the sparse high bit rate traffics from powerful devices?

A technological answer to the densification challenge is also embodied by long-range low-power networks such as SigFox, LoRa, NB-IoT, etc. In this context, the idea of offloading cellular traffic to different wireless access technologies is emerging as a very promising solution to relieve the traditional mobile network from its overwhelming load. In fact, offloading is already employed today, and, globally, 45% of total mobile data traffic was offloaded onto the fixed network through Wi-Fi or femtocells in 2013. Device-to-device (D2D) communications in hybrid networks, combining long-range cellular links and short-range technologies, opens even more possibilities. We aim at providing solutions that are missing for efficiently and practically mix multi-hop and cellular networks technologies.

Cellular M2M. Enabling a communication in a cellular network follows two major procedures: a resource allocation demand is first transmitted by the UE which, if successful, is followed by the actual data transmission phase, using dedicated resources allocated by the eNodeB (eNB) to the UE. This procedure was designed specifically for H2H traffic, which is bursty by nature, and it is based on the notions of session and call, activities that keep the user involved for a relatively long time and necessitate the exchange of a series of messages with the network. On the contrary, M2M traffic generates low amounts of data periodically or sporadically. Going through a signaling-heavy random access (RA) procedure to transmit one short message is strongly inefficient for both the M2M devices and the infrastructure.

In the perspective of 5G solutions, we are investigating mechanisms that regulate the M2M traffics in order to obtain good performances while keeping a reasonable quality of service (QoS) for human-to-human (H2H) terminals. The idea of piggybacking the M2M data transmission within one of the RA procedure messages is tempting and it is now considered as the best solution for this type of traffic. This means that the M2M data is transmitted on the shared resources of the RACH, and raises questions regarding the capacity of the RACH, which was not designed for these purposes. In this regard, our analysis of the access capacity of LTE-A

RACH procedure has to be adapted to multi-class scenarios, in order to understand the competition between M2M and H2H devices. Modeling based on Markov chains provides trends on system scale performances, while event-based simulations enable the analysis of the distribution of the performances over the different kinds of users. Combining both should give enough insights so as to design relevant regulation techniques and strategies. In particular two open questions that have to be tackled can be stated as: When should access resources be opened to M2M traffics without penalizing H2H performances? Does an eNodeB have a detailed enough knowledge of the system and transmit enough information to UE to regulate the traffics? The objective is to go to the analysis of achievable performances to actual protocols that take into account realistic M2M traffic patterns.

Hybrid networks. The first objective in this research axis is a realistic large-scale performance evaluation of Wi-Fi offloading solutions. While the mechanisms behind Wi-Fi offloading are now clear in the research community, their performance has only been tested in small-scale field tests, covering either small geographical areas (i.e. a few cellular base stations) and/or a small number of specific users (e.g. vehicular users). Instead, we evaluate the offloading performance at a city scale, building on real mobile network traces available in the team. First of all, through our collaboration with Orange Labs, we have access to an accurate characterization of the mobile traffic load at each base station in all major French cities. Second, a data collection application for Android devices has been developed in the team and used by hundreds of users in the Lyon metropolitan area. This application monitors and logs all the Wi-Fi access points in the coverage range of the smartphone, allowing us to build a map of Wi-Fi accessibility in some parts of the city. Combining these two data sources and completing them with simulation studies will allow an accurate evaluation of Wi-Fi offloading solutions over a large area.

On the D2D side, our focus is on the connected objects scenario, where we study the integration of short-range links and long-range technologies such as LTE, SigFox or LoRa. This requires the design of network protocols to discover and group the devices in a certain region. For this, we build on our expertise on clustering sensor and vehicular nodes. The important difference in this case is that the cellular network can assist the clustering formation process. The next step is represented by the selection of the devices that will be using the long-range links on behalf of the entire cluster. With respect to classical cluster head selection problems in ad-hoc networks, our problem distinguishes itself through device heterogeneity in terms of available communication technologies (not all devices have a long-range connection, or their quality is poor), energy resources (some devices might have energy harvesting capabilities) and expected lifetime. We will evaluate the proposed mechanisms both analytically (clustering problems are generally modeled by dominating set problems in graph theory) and through discrete-event simulation. Prototyping and experimental evaluation in cooperation with our industrial partners is also foreseen in this case.

3.3. Network data exploitation

Mobile devices are continuously interacting with the network infrastructure, and the associated geo-referenced events can be easily logged by the operators, for different purposes, including billing and resource management. This leads to the implicit possibility of monitoring a large percentage of the whole population with minimal cost: no other technology provides today an equivalent coverage. On the networking side, the exploration of data collected within the cellular network can be the enabler of flexible and reconfigurable cellular systems. In order to enable this vision, algorithmic solutions are needed that drive, in concert with the variations in the mobile demand, the establishment, modification, release and relocation of any type of resources in the network. This raises, in turn, the fundamental problem of understanding the mobile demand, and linking it to the resource management processes. More precisely, we contribute to answer questions about the correlation between urban areas and mobile traffic usage, in particular the spatial and temporal causalities in the usage of the mobile network.

In a different type of architecture, the one of wireless sensor networks, the spatio-temporal characteristics of the data that are transported can also be leveraged to improve on the networking performances, e.g. capacity and energy consumption. In several applications (e.g. temperature monitoring, intrusion detection), wireless sensor nodes are prone to transmit redundant or correlated information. This wastes the bandwidth

and accelerates the battery depletion. Energy and network capacity savings can be obtained by leveraging spatial and temporal correlation in packet aggregation. Packet transmissions can be reduced with an overhead induced by distributed aggregation algorithms. We aim at designing data aggregation functions that preserve data accuracy and maximize the network lifetime with low assumptions on the network topology and the application.

Mobile data analysis. In this research axis, we delve deeper in the analysis of mobile traffic. In this sense, temporal and spatial usage profiles can be built, by including in our analysis datasets providing service-level usage information. Indeed, previous studies have been generally using call detail records (CDR) or, at best, aggregated packet traffic information. This data is already very useful in many research fields, but fine-grained usage data would allow an even better understanding of the spatiotemporal characteristics of mobile traffic. To achieve this, we exploit datasets made available by Orange Labs, providing information about the network usage for several different mobile services (web, streaming, download, mail, etc.).

To obtain even richer information, we combine this operator-side data with user-side data, collected by a crowdsensing application we developed within the PrivaMov research project. While covering hundreds of thousands of users, operator data only allows to localize the user at the cell level, and only when the user is connected to the network. The crowdsensing application we are using gathers precise GPS user localization data at a high frequency. Combining these two sources of data will allow us to gain insight in possible biases introduced by operator-side data and to infer microscopic properties which, correctly modeled, can be extended to the entire user population, even those for which we do not possess crowdsensed data.

Privacy preservation is an important topic in the field of mobile data analysis. Mobile traffic data anonymization techniques are currently proposed, mainly by adding noise or removing information from the original dataset. While we do not plan to develop anonymization algorithms, we collaborate with teams working on this topic (e.g. Inria Privatics) in order to assess the impact of anonymization techniques on the spatio-temporal properties of mobile traffic data. Through a statistical analysis of both anonymized and non-anonymized data, we hope to better understand the usability of anonymized data for different applications based on the exploration of mobile traffic data.

Data aggregation. Data-aggregation takes benefit from spatial and/or temporal correlation, while preserving the data accuracy. Such correlation comes from the physical phenomenon which is observed. Temporal aggregation is mainly addressed using temporal series (e.g. ARMA) whereas spatial aggregation is now led by compressive sensing solutions. Our objective is to get rid of the assumption of knowing of the network topology properties and the data traffic generated by the application, in particular for dense and massive wireless networks. Note that we focus on data-aggregation with a networking perspective, not with the background of information theory.

The rational design of an aggregation scheme implies understanding data dynamics (statistical characteristics, information representation), algorithmic optimization (aggregator location, minimizing the number of aggregators toward energy efficiency), and network dynamics (routing, medium sharing policies, node activity). We look for designing a complete aggregation chain including both intra-sensor aggregation and inter-sensor aggregation. For this, we characterize the raw data that are collected in order to understand the dynamics behind several key applications. The goal is to provide a taxonomy of the applications according to the data properties in terms of stationarity, dynamics, etc. Then, we aim to design temporal aggregation functions without knowledge of the network topology and without assumptions about the application data. Such functions should be able to self-adapt to the environment evolution. A related issue is the deployment of aggregators into the wireless network to allow spatial aggregation with respect to the energy consumption minimization, capacity saving maximization and distributed algorithm complexity. We therefore look to define dedicated protocols for each aggregation function family.

ALPINES Project-Team

3. Research Program

3.1. Overview

The research described here is directly relevant to several steps of the numerical simulation chain. Given a numerical simulation that was expressed as a set of differential equations, our research focuses on mesh generation methods for parallel computation, novel numerical algorithms for linear algebra, as well as algorithms and tools for their efficient and scalable implementation on high performance computers. The validation and the exploitation of the results is performed with collaborators from applications and is based on the usage of existing tools. In summary, the topics studied in our group are the following:

- Numerical methods and algorithms
  - Mesh generation for parallel computation
  - Solvers for numerical linear algebra
    - Domain decomposition methods
    - Preconditioning for iterative methods
  - Computational kernels for numerical linear algebra
  - Tensor computations
- Validation on numerical simulations and other numerical applications

3.2. Domain specific language - parallel FreeFem++

In the engineering, researchers, and teachers communities, there is a strong demand for simulation frameworks that are simple to install and use, efficient, sustainable, and that solve efficiently and accurately complex problems for which there are no dedicated tools or codes available. In our group we develop FreeFem++ (see https://www.freefem.org/), a user dedicated language for solving PDEs. The goal of FreeFem++ is not to be a substitute for complex numerical codes, but rather to provide an efficient and relatively generic tool for:

- getting a quick answer to a specific problem,
- prototyping the resolution of a new complex problem.

The current users of FreeFem++ are mathematicians, engineers, university professors, and students. In general for these users the installation of public libraries as MPI, MUMPS, Ipopt, Blas, lapack, OpenGL, fftw, scotch, PETSc, SLEPc is a very difficult problem. For this reason, the authors of FreeFem++ have created a user friendly language, and over years have enriched its capabilities and provided tools for compiling FreeFem++ such that the users do not need to have special knowledge of computer science. This leads to an important work on porting the software on different emerging architectures.

Today, the main components of parallel FreeFem++ are:

1. definition of a coarse grid,
2. splitting of the coarse grid,
3. mesh generation of all subdomains of the coarse grid, and construction of parallel data structures for vectors and sparse matrices from the mesh of the subdomain,
4. call to a linear solver,
5. analysis of the result.

All these components are parallel, except for point (5) which is not in the focus of our research. However for the moment, the parallel mesh generation algorithm is very simple and not sufficient, for example it addresses only polygonal geometries. Having a better parallel mesh generation algorithm is one of the goals of our project. In addition, in the current version of FreeFem++, the parallelism is not hidden from the user, it is done through direct calls to MPI. Our goal is also to hide all the MPI calls in the specific language part of FreeFem++. In addition to these in-house domain decomposition methods, FreeFem++ is also linked to PETSc solvers which enables an easy use of third parties parallel multigrid methods.

3.3. Solvers for numerical linear algebra

Iterative methods are widely used in industrial applications, and preconditioning is the most important research subject here. Our research considers domain decomposition methods and iterative methods and its goal is to develop solvers that are suitable for parallelism and that exploit the fact that the matrices are arising from the discretization of a system of PDEs on unstructured grids.

One of the main challenges that we address is the lack of robustness and scalability of existing methods as incomplete LU factorizations or Schwarz-based approaches, for which the number of iterations increases significantly with the problem size or with the number of processors. This is often due to the presence of several low frequency modes that hinder the convergence of the iterative method. To address this problem, we study different approaches for dealing with the low frequency modes as coarse space correction in domain decomposition or deflation techniques.

We also focus on developing boundary integral equation methods that would be adapted to the simulation of wave propagation in complex physical situations, and that would lend themselves to the use of parallel architectures. The final objective is to bring the state of the art on boundary integral equations closer to contemporary industrial needs. From this perspective, we investigate domain decomposition strategies in conjunction with boundary element method as well as acceleration techniques (H-matrices, FMM and the like) that would appear relevant in multi-material and/or multi-domain configurations. Our work on this topic also includes numerical implementation on large scale problems, which appears as a challenge due to the peculiarities of boundary integral equations.

3.4. Computational kernels for numerical linear and multilinear algebra

The design of new numerical methods that are robust and that have well proven convergence properties is one of the challenges addressed in Alpines. Another important challenge is the design of parallel algorithms for the novel numerical methods and the underlying building blocks from numerical linear algebra. The goal is to enable their efficient execution on a diverse set of node architectures and their scaling to emerging high-performance clusters with an increasing number of nodes.

Increased communication cost is one of the main challenges in high performance computing that we address in our research by investigating algorithms that minimize communication, as communication avoiding algorithms. We propose to integrate the minimization of communication into the algorithmic design of numerical linear algebra problems. This is different from previous approaches where the communication problem was addressed as a scheduling or as a tuning problem. The communication avoiding algorithmic design is an approach originally developed in our group since 2007 (initially in collaboration with researchers from UC Berkeley and CU Denver). While at mid term we focus on reducing communication in numerical linear algebra, at long term we aim at considering the communication problem one level higher, during the parallel mesh generation tool described earlier.

Our research also focuses on solving problems of large size that feature high dimensions as in molecular simulations. The data in this case is represented by objects called tensors, or multilinear arrays. The goal is to design novel tensor techniques to allow their effective compression, i.e. their representation by simpler objects in small dimensions, while controlling the loss of information. The algorithms are aiming to being highly parallel to allow to deal with the large number of dimensions and large data sets, while preserving the required information for obtaining the solution of the problem.

AVALON Project-Team

3. Research Program

3.1. Energy Application Profiling and Modeling

Despite recent improvements, there is still a long road to follow in order to obtain energy efficient, energy proportional and eco-responsible exascale systems by 2022. Energy efficiency is therefore a major challenge for building next generation large-scale platforms. The targeted platforms will gather hundreds of millions of cores, low power servers, or CPUs. Besides being very important, their power consumption will be dynamic and irregular.

Thus, to consume energy efficiently, we aim at investigating two research directions. First, we need to improve measurement, understanding, and analysis on how large-scale platforms consume energy. Unlike some approaches [24] that mix the usage of internal and external wattmeters on a small set of resources, we target high frequency and precise internal and external energy measurements of each physical and virtual resource on large-scale distributed systems.

Secondly, we need to find new mechanisms that consume less and better on such platforms. Combined with hardware optimizations, several works based on shutdown or slowdown approaches aim at reducing energy consumption of distributed platforms and applications. To consume less, we first plan to explore the provision of accurate estimation of the energy consumed by applications without pre-executing and knowing them while most of the works try to do it based on in-depth application knowledge (code instrumentation [27], phase detection for specific HPC applications [31], *etc.*). As a second step, we aim at designing a framework model that allows interaction, dialogue and decisions taken in cooperation among the user/application, the administrator, the resource manager, and the energy supplier. While smart grid is one of the last killer scenarios for networks, electrical provisioning of next generation large IT infrastructures remains a challenge.

3.2. Data-intensive Application Profiling, Modeling, and Management

Recently, the term “Big Data” has emerged to design data sets or collections so large that they become intractable for classical tools. This term is most time implicitly linked to “analytics” to refer to issues such as data curation, storage, search, sharing, analysis, and visualization. However, the Big Data challenge is not limited to data-analytics, a field that is well covered by programming languages and run-time systems such as Map-Reduce. It also encompasses data-intensive applications. These applications can be sorted into two categories. In High Performance Computing (HPC), data-intensive applications leverage post-petascale infrastructures to perform highly parallel computations on large amount of data, while in High Throughput Computing (HTC), a large amount of independent and sequential computations are performed on huge data collections.

These two types of data-intensive applications (HTC and HPC) raise challenges related to profiling and modeling that the AVALON team proposes to address. While the characteristics of data-intensive applications are very different, our work will remain coherent and focused. Indeed, a common goal will be to acquire a better understanding of both the applications and the underlying infrastructures running them to propose the best match between application requirements and infrastructure capacities. To achieve this objective, we will extensively rely on logging and profiling in order to design sound, accurate, and validated models. Then, the proposed models will be integrated and consolidated within a single simulation framework (SIMGRID). This will allow us to explore various potential “what-if?” scenarios and offer objective indicators to select interesting infrastructure configurations that match application specificities.

Another challenge is the ability to mix several heterogeneous infrastructures that scientists have at their disposal (*e.g.,* Grids, Clouds, and Desktop Grids) to execute data-intensive applications. Leveraging the aforementioned results, we will design strategies for efficient data management service for hybrid computing infrastructures.

3.3. Resource-Agnostic Application Description Model

With parallel programming, users expect to obtain performance improvement, regardless its cost. For long, parallel machines have been simple enough to let a user program use them given a minimal abstraction of their hardware. For example, MPI [26] exposes the number of nodes but hides the complexity of network topology behind a set of collective operations; OpenMP [30] simplifies the management of threads on top of a shared memory machine while OpenACC [29] aims at simplifying the use of GPGPU.

However, machines and applications are getting more and more complex so that the cost of manually handling an application is becoming very high [25]. Hardware complexity also stems from the unclear path towards next generations of hardware coming from the frequency wall: multi-core CPU, many-core CPU, GPGPUs, deep memory hierarchy, *etc.* have a strong impact on parallel algorithms. Parallel languages (UPC, Fortress, X10, *etc.*) is a first piece of the solution. However, they will still face the challenge of supporting distinct codes corresponding to different algorithms corresponding to distinct hardware capacities.

Therefore, the challenge we aim to address is to define a model, for describing the structure of parallel and distributed applications that enables code variations but also efficient executions on parallel and distributed infrastructures. Indeed, this issue appears for HPC applications but also for cloud oriented applications. The challenge is to adapt an application to user constraints such as performance, energy, security, *etc.*

Our approach is to consider component based models [32] as they offer the ability to manipulate the software architecture of an application. To achieve our goal, we consider a "compilation" approach that transforms a resource-agnostic application description into a resource-specific description. The challenge is thus to determine a component based model that enables to efficiently compute application mapping while being tractable. In particular, it has to provide an efficient support with respect to application and resource elasticity, energy consumption and data management. OpenMP runtime is a specific use case that we target.

3.4. Application Mapping and Scheduling

This research axis is at the crossroad of the AVALON team. In particular, it gathers results of the three other research axis. We plan to consider application mapping and scheduling addressing the following three issues.

3.4.1. Application Mapping and Software Deployment

Application mapping and software deployment consist in the process of assigning distributed pieces of software to a set of resources. Resources can be selected according to different criteria such as performance, cost, energy consumption, security management, *etc.* A first issue is to select resources at application launch time. With the wide adoption of elastic platforms, *i.e.,* platforms that let the number of resources allocated to an application to be increased or decreased during its execution, the issue is also to handle resource selection at runtime.

The challenge in this context corresponds to the mapping of applications onto distributed resources. It will consist in designing algorithms that in particular take into consideration application profiling, modeling, and description.

A particular facet of this challenge is to propose scheduling algorithms for dynamic and elastic platforms. As the number of elements can vary, some kind of control of the platforms must be used accordingly to the scheduling.

3.4.2. Non-Deterministic Workflow Scheduling

Many scientific applications are described through workflow structures. Due to the increasing level of parallelism offered by modern computing infrastructures, workflow applications now have to be composed not only of sequential programs, but also of parallel ones. New applications are now built upon workflows with conditionals and loops (also called non-deterministic workflows).

These workflows cannot be scheduled beforehand. Moreover cloud platforms bring on-demand resource provisioning and pay-as-you-go billing models. Therefore, there is a problem of resource allocation for non-deterministic workflows under budget constraints and using such an elastic management of resources.

Another important issue is data management. We need to schedule the data movements and replications while taking job scheduling into account. If possible, data management and job scheduling should be done at the same time in a closely coupled interaction.

3.4.3. Software Asset Management

The use of software is generally regulated by licenses, whether they are free or paid and with or without access to their sources. The world of licenses is very vast and unknown (especially in the industrial world). Often only the general public version is known (a software purchase corresponds to a license). For enterprises, the reality is much more complex, especially for main publishers. We work on the OpTISAM software, a software offering tools to perform Software Asset Management (SAM) much more efficiently in order to be able to ensure the full compliance with all contracts from each software and a new type of deployment taking into account these aspects and other additional parameters like energy and performance. This work is built on an Orange™ collaboration.

3.4.4. Cloud deployment and reproducibility

As part of the scientific method, any researcher should be able to reproduce the experimentation in order to not only verify the result but also evaluate and compare this experimentation with other approaches. The need of a standard tool allowing researchers to easily generate, share and reproduce experiments set-up arises. In our research, through a Nokia collaboration, we created SeeDep [10], a framework aiming at being such a standard tool. By associating a generation key to a network experiment set-up, SeeDep allows for reproducing network experiments independently from the used infrastructure.

COAST Project-Team

3. Research Program

3.1. Introduction

Our scientific foundations are grounded on distributed collaborative systems supported by sophisticated data sharing mechanisms and on service oriented computing with an emphasis on orchestration and on non-functional properties. Distributed collaborative systems enable distributed group work supported by computer technologies. Designing such systems requires an expertise in Distributed Systems and in Computer-supported collaborative Work research area. Besides theoretical and technical aspects of distributed systems, the design of distributed collaborative systems must take into account the human factor to offer solutions suitable for users and groups. The Coast team vision is to move away from a centralized authority based collaboration toward a decentralized collaboration. Users will have full control over their data. They can store them locally and decide with whom to share them. The Coast team investigates the issues related to the management of distributed shared data and coordination between users and groups. Service oriented Computing [16] is an established domain on which the ECOO, Score and now the Coast teams have been contributing for a long time. It refers to the general discipline that studies the development of computer applications on the web. A service is an independent software program with a specific functional context and capabilities published as a service contract (or more traditionally an API). A service composition aggregates a set of services and coordinates their interactions. The scale, the autonomy of services, the heterogeneity and some design principles underlying Service Oriented Computing open new research questions that are at the basis of our research. They span the disciplines of **distributed computing**, **software engineering** and **computer supported collaborative work** (CSCW). Our approach to contribute to the general vision of Service Oriented Computing is to focus on the issue of the efficient and flexible construction of reliable and secure high-level services. We aim to achieve it through the coordination/orchestration/composition of other services provided by distributed organizations or people.

3.2. Consistency Models for Distributed Collaborative Systems

Collaborative systems are distributed systems that allow users to share data. One important issue is to manage consistency of shared data according to concurrent access. Traditional consistency criteria such as serializability, linearizability are not adequate for collaborative systems. Causality, Convergence and Intention preservation (CCI) [21] are more suitable for developing middleware for collaborative applications. We develop algorithms for ensuring CCI properties on collaborative distributed systems. Constraints on the algorithms are different according to the kind of distributed system and to the data structure. The distributed system can be centralized, decentralized or peer-to-peer. The type of data can include strings, growable arrays, ordered trees, semantic graphs and multimedia data.

3.3. Optimistic Replication

Replication of data among different nodes of a network promotes reliability, fault tolerance, and availability. When data are mutable, consistency among the different replicas must be ensured. Pessimistic replication is based on the principle of single-copy consistency while optimistic replication allows the replicas to diverge during a short time period. The consistency model for optimistic replication [19] is called eventual consistency, meaning that replicas are guaranteed to converge to the same value when the system is idle. Our research focuses on the two most promising families of optimistic replication algorithms for ensuring CCI:

- operational transformation (OT) algorithms [14]
- algorithms based on commutative replicated data types (CRDT) [18].

Operational transformation algorithms are based on the application of a transformation function when a remote modification is integrated into the local document. Integration algorithms are generic, being parametrised by operational transformation functions which depend on replicated document types. The advantage of these algorithms is their genericity. These algorithms can be applied to any data type and they can merge heterogeneous data in a uniform manner. Commutative replicated data types is a new class of algorithms initiated by WooT [15], the first algorithm designed WithOut Operational Transformations. They ensure consistency of highly dynamic content on peer-to-peer networks. Unlike traditional optimistic replication algorithms, they can ensure consistency without concurrency control. CRDT algorithms rely on natively commutative operations defined on abstract data types such as lists or ordered trees. Thus, they do not require a merge algorithm or an integration procedure.

3.4. Process Orchestration and Management

Process Orchestration and Management is considered as a core discipline behind Service Management and Computing. It includes the analysis, the modelling, the execution, the monitoring and the continuous improvement of enterprise processes and is for us a central domain of studies. Many efforts have been devoted establishing standard business process models founded on well-grounded theories (e.g. Petri Nets) that meet the needs of business analysts, software engineers and software integrator. This led to heated debate in the Business Process Management (BPM) community as the two points of view are very difficult to reconcile. On one side, business people in general require models that are easy to use and understand and that can be quickly adapted to exceptional situations. On the other side, IT people need models with an operational semantic in order to be able transform them into executable artifacts. Part of our work has been an attempt to reconcile these points of view. This resulted in the development of the Bonita BPM system. It resulted also more recently on our work in crisis management where the same people are designing, executing and monitoring the process as it executes. More generally, and at a larger scale, we have been considering the problem of processes spanning the barriers of organizations. This leads to the more general problem of service composition as a way to coordinate inter organizational construction of applications. These applications provide value, based on the composition of lower level services [12].

3.5. Service Composition

Recently, we started a study on service composition for software architects where services are coming from different providers with different plans (capacity, degree of resilience...). The objective is to support the architects to select the most accurate services (wrt. to their requirements, both functional and non-functional) and plans for building their software. We also compute the properties that we enforce for the composition of these services.

COATI Project-Team

3. Research Program

3.1. Research Program

Members of COATI have a strong expertise in the design and management of wired and wireless backbone, backhaul, broadband, software defined and complex networks. On the one hand, we cope with specific problems such as energy efficiency in backhaul and backbone networks, routing reconfiguration in connection oriented networks (MPLS, WDM), traffic aggregation in SONET networks, compact routing in large-scale networks, survivability to single and multiple failures, etc. These specific problems often come from questions of our industrial partners. On the other hand, we study fundamental problems mainly related to routing and reliability that appear in many networks (not restricted to our main fields of applications) and that have been widely studied in the past. However, previous solutions do not take into account the constraints of current networks/traffic such as their huge size and their dynamics. COATI thus puts a significant research effort in the following directions:

- **Service Function Chains (SFC):** we study the placement of Service Function Chains within the network considering the ordering constraints. Then, we focus firstly on energy efficiency and secondly on reliability and protection mechanisms. In a last step, we study reconfiguration of the SFCs in case of dynamic traffic with a make-before-break approach.
- **Larger networks:** Another challenge one has to face is the increase in size of practical instances. It is already difficult, if not impossible, to solve practical instances optimally using existing tools. Therefore, we have to find new ways to solve problems using reduction and decomposition methods, characterization of polynomial instances (which are surprisingly often the practical ones), or algorithms with acceptable practical performances.
- **Stochastic behaviors:** Larger topologies mean frequent changes due to traffic and radio fluctuations, failures, maintenance operations, growth, routing policy changes, etc. We aim at including these stochastic behaviors in our combinatorial optimization process to handle the dynamics of the system and to obtain robust designs of networks.

The methods and tools used in our studies come from discrete mathematics and combinatorial optimization, and COATI contributes to their improvements. Also, COATI works on graph-decomposition methods and various games on graphs which are essential for a better understanding of the structural and combinatorial properties of the problems, but also for the design of efficient exact or approximate algorithms. We contribute to the modelling of optimization problems in terms of graphs, study the complexity of the problems, and then we investigate the structural or metric properties of graphs that make these problems hard or easy. We exploit these properties in the design of algorithms in order to find the most efficient ways for solving the problems.

COATI also focuses on the theory of *directed graphs*. Indeed, graph theory can be roughly partitioned into two branches: the areas of undirected graphs and directed graphs. Even though both areas have numerous important applications, for various reasons, undirected graphs have been studied much more extensively than directed graphs. It is worth noticing that many telecommunication problems are modelled with directed graphs. Therefore, a deeper understanding of the theory of directed graphs will benefit to the resolution of telecommunication networks problems. For instance, the problem of finding disjoint paths becomes much more difficult in directed graphs and understanding the underlying structures of actual directed networks would help us to propose solutions.

Last, we have recently started investigating how tools from multi-agents based systems and machine learning theory could help solving some optimization problems in networks. The arrival of Emanuele Natale as a Junior Researcher (CNRS) in the team and of two new PhD students (Franceso D'Amore and Hicham Lesfari) will foster these investigations.

CTRL-A Project-Team

3. Research Program

3.1. Modeling and control techniques for autonomic computing

The main objective of CTRL-A translates into a number of scientific challenges, the most important of these are:

- (i) programming language support, on the two facets of model-oriented languages, based on automata [5], and of domain specific languages, following e.g., a component-based approach [4], [1] or related to rule-based or HMI languages ;
- (ii) design methods for reconfiguration controller design in computing systems, proposing generic systems architectures and models based on transition systems [3], classical continuous control or controlled stochastic systems.

  We adopt a strategy of constant experimental identification of needs and validation of proposals, in application domains like middleware platforms for Cloud systems [3], multi-core HPC architectures [10], Dynamic Partial Reconfiguration in FPGA-based hardware [2] and the IoT and smart environments [8].

Achieving the goals of CTRL-A requires multidisciplinarity and expertise from several domains. The expertise in Autonomic Computing and programming languages is covered internally by members of the Ctrl-A team. On the side of theoretical aspects of control, we have active external collaborations with researchers specialized in Control Theory, in the domain of Discrete Event Systems as well as in classical, continuous control. Additionally, an important requirement for our research to have impact is to have access to concrete, real-world computing systems requiring reconfiguration control. We target autonomic computing at different scales, in embedded systems or in cloud infrastructures, which are traditionally different domains. This is addressed by external collaborations, with experts in either hardware or software platforms, who are generally missing our competences on model-based control of reconfigurations.

DANTE Project-Team

3. Research Program

3.1. Graph-based signal processing

Participants: Paulo Gonçalves, Rémi Gribonval, Marion Foare, Márton Karsai.

Evolving networks can be regarded as "*out of equilibrium*" systems. Indeed, their dynamics are typically characterized by non standard and intricate statistical properties, such as non-stationarity, long range memory effects, intricate space and time correlations.

Analyzing, modeling, and even defining adapted concepts for dynamic graphs is at the heart of DANTE. This is a largely open question that has to be answered by keeping a balance between specificity (solutions triggered by specific data sets) and generality (universal approaches disconnected from social realities). We will tackle this challenge from a graph-based signal processing perspective involving signal analysts and computer scientists, together with experts of the data domain application. One can distinguish two different issues in this challenge, one related to the graph-based organization of the data and the other to the time dependency that naturally exits in the dynamic graph object. In both cases, a number of contributions can be found in the literature, albeit in different contexts. In our application domain, high-dimensional data "naturally reside" on the vertices of weighted graphs. The emerging field of signal processing on graphs merges algebraic and spectral graph theoretic concepts with computational harmonic analysis to process such signals on graphs [74].

As for the first point, adapting well-founded signal processing techniques to data represented as graphs is an emerging, yet quickly developing field which has already received key contributions. Some of them are very general and delineate ambitious programs aimed at defining universal, generally unsupervised methods for exploring high-dimensional data sets and processing them. This is the case for instance of the “diffusion wavelets” and “diffusion maps” pushed forward at Yale and Duke [57]. Others are more traditionally connected with standard signal processing concepts, in the spirit of elaborating new methodologies via some bridging between networks and time series, see for instance [69] and references therein. Other viewpoints can be found as well, including multi-resolution Markov models [77], Bayesian networks or distributed processing over sensor networks [68]. Such approaches can be particularly successful for handling static graphs and unveiling aspects of their organization in terms of dependencies between nodes, grouping, etc. Incorporating possible time dependencies within the whole picture calls however for the addition of an extra dimension to the problem "as it would be the case when switching from one image to a video sequence", a situation for which one can imagine to take advantage of the whole body of knowledge attached to non-stationary signal processing [58].

The arrival of Rémi Gribonval in August 2019 brought a new dimension to the research program of this theme. Specialist of parsimonious representations of large data sets, R. Gribonval will develop at Dante a specific activity related to the sparsification of resources (computing and storage but also regarding the data volumes) in the context of machine and deep learning. This new orientation of Dante will be elaborated and fully integrated to the objectives of the future Inria project that will be proposed after Dante.

3.2. Theory and Structure of dynamic Networks

Participants: Christophe Crespelle, Anthony Busson, Márton Karsai, Éric Guichard.

Characterization of the dynamics of complex networks. We need to focus on intrinsic properties of evolving/dynamic complex networks. New notions (as opposed to classical static graph properties) have to be introduced: rate of vertices or links appearances or disappearances, the duration of link presences or absences. Moreover, more specific properties related to the dynamics have to be defined and are somehow related to the way to model a dynamic graph.

Through the systematic analysis and characterization of static network representations of many different systems, researchers of several disciplines have unveiled complex topologies and heterogeneous structures, with connectivity patterns statistically characterized by heavy-tails and large fluctuations, scale-free properties and non trivial correlations such as high clustering and hierarchical ordering [71]. A large amount of work has been devoted to the development of new tools for statistical characterisation and modelling of networks, in order to identify their most relevant properties, and to understand which growth mechanisms could lead to these properties. Most of those contributions have focused on static graphs or on dynamic process (*e.g.* diffusion) occurring on static graphs. This has called forth a major effort in developing the methodology to characterize the topology and temporal behaviour of complex networks [71], [62], [78], [67], to describe the observed structural and temporal heterogeneities [55], [62], [56], to detect and measure emerging community structures [59], [75], [76], to see how the functionality of networks determines their evolving structure [66], and to determine what kinds of correlations play a role in their dynamics [63], [65], [70].

The challenge is now to extend this kind of statistical characterization to dynamical graphs. In other words, links in dynamic networks are temporal events, called contacts, which can be either punctual or last for some period of time. Because of the complexity of this analysis, the temporal dimension of the network is often ignored or only roughly considered. Therefore, fully taking into account the dynamics of the links into a network is a crucial and highly challenging issue.

Another powerful approach to model time-varying graphs is via activity driven network models. In this case, the only assumption relates to the distribution of activity rates of interacting entities. The activity rate is realistically broadly distributed and refers to the probability that an entity becomes active and creates a connection with another entity within a unit time step [73]. Even the generic model is already capable to recover some realistic features of the emerging graph, its main advantage is to provide a general framework to study various types of correlations present in real temporal networks. By synthesising such correlations (*e.g.* memory effects, preferential attachment, triangular closing mechanisms, ...) from the real data, we are able to extend the general mechanism and build a temporal network model, which shows certain realistic feature in a controlled way. This can be used to study the effect of selected correlations on the evolution of the emerging structure [64] and its co-evolution with ongoing processes like spreading phenomena, synchronisation, evolution of consensus, random walk etc. [64], [72]. This approach allows also to develop control and immunisation strategies by fully considering the temporal nature of the backgrounding network.

3.3. Distributed Algorithms for dynamic networks: regulation, adaptation and interaction

Participants: Thomas Begin, Anthony Busson, Isabelle Guérin Lassous, Philippe Nain.

Dedicated algorithms for dynamic networks. First, the dynamic network object itself trigger original algorithmic questions. It mainly concerns distributed algorithms that should be designed and deployed to efficiently measure the object itself and get an accurate view of its dynamic behavior. Such distributed measure should be “transparent”, that is, it should introduce no bias or at least a bias that is controllable and corrigible. Such problem is encountered in all distributed metrology measures / distributed probes: P2P, sensor network, wireless network, QoS routing... This question raises naturally the intrinsic notion of adaptation and control of the dynamic network itself since it appears that autonomous networks and traffic aware routing are becoming crucial.

Communication networks are dynamic networks that potentially undergo high dynamicity. The dynamicity exhibited by these networks results from several factors including, for instance, changes in the topology and varying workload conditions. Although most implemented protocols and existing solutions in the literature can cope with a dynamic behavior, the evolution of their behavior operates identically whatever the actual properties of the dynamicity. For instance, parameters of the routing protocols (*e.g.* hello packets transmission frequency) or routing methods (*e.g.* reactive / proactive) are commonly hold constant regardless of the nodes mobility. Similarly, the algorithms ruling CSMA/CA (*e.g.* size of the contention window) are tuned identically and they do not change according to the actual workload and observed topology.

Dynamicity in computer networks tends to affect a large number of performance parameters (if not all) coming from various layers (viz. physical, link, routing and transport). To find out which ones matter the most for our intended purpose, we expect to rely on the tools developed by the two former axes. These quantities should capture and characterize the actual network dynamicity. Our goal is to take advantage of this latter information in order to refine existing protocols, or even to propose new solutions. More precisely, we will attempt to associate "fundamental" changes occurring in the underlying graph of a network (reported through graph-based signal tools) to quantitative performance that are matter of interests for networking applications and the end-users. We expect to rely on available testbeds such as SensLab and FIT to experiment our solutions and ultimately validate our approach.

DATAMOVE Project-Team

3. Research Program

3.1. Motivation

Today's largest supercomputers [0] are composed of few millions of cores, with performances almost reaching 100 PetaFlops [0] for the largest machine. Moving data in such large supercomputers is becoming a major performance bottleneck, and the situation is expected to worsen even more at exascale and beyond. The data transfer capabilities are growing at a slower rate than processing power ones. The profusion of available flops will very likely be underused due to constrained communication capabilities. It is commonly admitted that data movements account for 50% to 70% of the global power consumption [0]. Thus, data movements are potentially one of the most important source of savings for enabling supercomputers to stay in the commonly adopted energy barrier of 20 MegaWatts. In the mid to long term, non volatile memory (NVRAM) is expected to deeply change the machine I/Os. Data distribution will shift from disk arrays with an access time often considered as uniform, towards permanent storage capabilities at each node of the machine, making data locality an even more prevalent paradigm.

The proposed DataMove team will work on **optimizing data movements for large scale computing** mainly at two related levels:

- Resource allocation
- Integration of numerical simulation and data analysis

The resource and job management system (also called batch scheduler or RJMS) is in charge of allocating resources upon user requests for executing their parallel applications. The growing cost of data movements requires adapted scheduling policies able to take into account the influence of intra-application communications, I/Os as well as contention caused by data traffic generated by other concurrent applications. Modelling the application behavior to anticipate its actual resource usage on such architecture is known to be challenging, but it becomes critical for improving performances (execution time, energy, or any other relevant objective). The job management system also needs to handle new types of workloads: high performance platforms now need to execute more and more often data intensive processing tasks like data analysis in addition to traditional computation intensive numerical simulations. In particular, the ever growing amount of data generated by numerical simulation calls for a tighter integration between the simulation and the data analysis. The challenge here is to reduce data traffic and to speed-up result analysis by performing result processing (compression, indexation, analysis, visualization, etc.) as closely as possible to the locus and time of data generation. This emerging trend called *in-situ analytics* requires to revisit the traditional workflow (loop of batch processing followed by postmortem analysis). The application becomes a whole including the simulation, in-situ processing and I/Os. This motivates the development of new well-adapted resource sharing strategies, data structures and parallel analytics schemes to efficiently interleave the different components of the application and globally improve the performance.

3.2. Strategy

DataMove targets HPC (High Performance Computing) at Exascale. But such machines and the associated applications are expected to be available only in 5 to 10 years. Meanwhile, we expect to see a growing number of petaflop machines to answer the needs for advanced numerical simulations. A sustainable exploitation of these petaflop machines is a real and hard challenge that we will address. We may also see in the coming years a convergence between HPC and Big Data, HPC platforms becoming more elastic and supporting Big

[0] Top500 Ranking, http://www.top500.org
[0] 10^{15} floating point operations per second
[0] SciDAC Review, 2010, http://scidacreview.org/1001/pdf/hardware.pdf

Data jobs, or HPC applications being more commonly executed on cloud like architectures. This is the second top objective of the 2015 US Strategic Computing Initiative [0]: *Increasing coherence between the technology base used for modelling and simulation and that used for data analytic computing*. We will contribute to that convergence at our level, considering more dynamic and versatile target platforms and types of workloads.

Our approaches should entail minimal modifications on the code of numerical simulations. Often large scale numerical simulations are complex domain specific codes with a long life span. We assume these codes as being sufficiently optimized. We will influence the behavior of numerical simulations through resource allocation at the job management system level or when interleaving them with analytics code.

To tackle these issues, we propose to intertwine theoretical research and practical developments in an agile mode. Algorithms with performance guarantees will be designed and experimented on large scale platforms with realistic usage scenarios developed with partner scientists or based on logs of the biggest available computing platforms (national supercomputers like Curie, or the BlueWaters machine accessible through our collaboration with Argonne National Lab). Conversely, a strong experimental expertise will enable to feed theoretical models with sound hypotheses, to twist proven algorithms with practical heuristics that could be further retro-feeded into adequate theoretical models.

A central scientific question is to make the relevant choices for optimizing performance (in a broad sense) in a reasonable time. HPC architectures and applications are increasingly complex systems (heterogeneity, dynamicity, uncertainties), which leads to consider the **optimization of resource allocation based on multiple objectives**, often contradictory (like energy and run-time for instance). Focusing on the optimization of one particular objective usually leads to worsen the others. The historical positioning of some members of the team who are specialists in multi-objective optimization is to generate a (limited) set of trade-off configurations, called *Pareto points*, and choose when required the most suitable trade-off between all the objectives. This methodology differs from the classical approaches, which simplify the problem into a single objective one (focus on a particular objective, combining the various objectives or agglomerate them). The real challenge is thus to combine algorithmic techniques to account for this diversity while guaranteeing a target efficiency for all the various objectives.

The DataMove team aims to elaborate generic and effective solutions of practical interest. We will make our new algorithms accessible through the team flagship software tools, **the OAR batch scheduler and the in-situ processing framework FlowVR**. We will maintain and enforce strong links with teams closely connected with large architecture design and operation (CEA DAM, BULL, Argonne National Lab), as well as scientists of other disciplines, in particular computational biologists, with whom we will elaborate and validate new usage scenarios (IBPC, CEA DAM, EDF).

3.3. Research Directions

DataMove research activity is organised around three directions. When a parallel job executes on a machine, it triggers data movements through the input data it needs to read, the results it produces (simulation results as well as traces) that need to be stored in the file system, as well as internal communications and temporary storage (for fault tolerance related data for instance). Modeling in details the simulation and the target machines to analyze scheduling policies is not feasible at large scales. We propose to investigate alternative approaches, including learning approaches, to capture and model the influence of data movements on the performance metrics of each job execution to develop **Data Aware Batch Scheduling** models and algorithms (Sec. 4.1). Experimenting new scheduling policies on real platforms at scale is unfeasible. Theoretical performance guarantees are not sufficient to ensure a new algorithm will actually perform as expected on a real platform. An intermediate evaluation level is required to probe novel scheduling policies. The second research axe focuses on the **Empirical Studies of Large Scale Platforms** (Sec. 4.2). The goal is to investigate how we could extract from actual computing centers traces information to replay the job allocations and executions on a simulated or emulated platform with new scheduling policies. Schedulers need information about jobs behavior on target machines to actually be able to make efficient allocation decisions. Asking users

[0] https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

to caracterize jobs often does not lead to reliable information. The third research direction **Integration of High Performance Computing and Data Analytics** (Sec. 4.3) addresses the data movement issue from a different perspective. New data analysis techniques on the HPC platform introduce new type of workloads, potentially more data than compute intensive, but could also enable to reduce data movements by directly enabling to pipe-line simulation execution with a live analysis of the produced results. Our goal is here to investigate how to program and schedule such analysis workflows in the HPC context.

DELYS Project-Team

3. Research Program

3.1. Research rationale

DELYS addresses both theoretical and practical issues of *Computer Systems*, leveraging our dual expertise in theoretical and experimental research. Our approach is a "virtuous cycle," triggered by issues with real systems, of algorithm design which we prove correct and evaluate theoretically, and then implement and test experimentally feeding back to theory. The major challenges addressed by DELYS are the sharing of information and guaranteeing correct execution of highly-dynamic computer systems. Our research covers a large spectrum of distributed computer systems: multicore computers, mobile networks, cloud computing systems, and dynamic communicating entities. This holistic approach enables handling related problems at different levels. Among such problems we can highlight consensus, fault detection, scalability, search of information, resource allocation, replication and consistency of shared data, dynamic content distribution, and concurrent and parallel algorithms.

Two main evolutions in the Computer Systems area strongly influence our research project:

(1) Modern computer systems are **increasingly distributed**, **dynamic** and composed of multiple devices **geographically spread over heterogeneous platforms**, spanning multiple management domains. Years of research in the field are now coming to fruition, and are being used by millions of users of web systems, peer-to-peer systems, gaming and social applications, cloud computing, and now fog computing. These new uses bring new challenges, such as *adaptation to dynamically-changing conditions*, where knowledge of the system state can only be partial and incomplete.

(2) **Heterogeneous architectures and virtualisation are everywhere**. The parallelism offered by distributed clusters and *multicore* architectures is opening highly parallel computing to new application areas. To be successful, however, many issues need to be addressed. Challenges include obtaining a consistent view of shared resources, such as memory, and optimally distributing computations among heterogeneous architectures. These issues arise at a more fine-grained level than before, leading to the need for different solutions down to OS level itself.

The scientific challenges of the distributed computing systems are subject to many important features which include scalability, fault tolerance, dynamics, emergent behaviour, heterogeneity, and virtualisation at many levels. Algorithms designed for traditional distributed systems, such as resource allocation, data storage and placement, and concurrent access to shared data, need to be redefined or revisited in order to work properly under the constraints of these new environments. Sometimes, classical "*static*" problems, (*e.g.*, Election Leader, Spanning Tree Construction, ...) even need to be redefined to consider the unstable nature of the distributed system. In particular, DELYS will focus on a number of key challenges:

Consistency in geo-scale systems. Distributed systems need to scale to large geographies and large numbers of attached devices, while executing in an untamed, unstable environment. This poses difficult scientific challenges, which are all the more pressing as the cloud moves more and more towards the edge, IoT and mobile computing. A key issue is how to share data effectively and consistently across the whole spectrum. Delys has made several key contributions, including CRDTs, the Transactional Causal Consistency Plus model, the AntidoteDB geo-distributed database, and its edge extension EdgeAnt.

Rethinking distributed algorithms. From a theoretical point of view the key question is how to adapt the fundamental building blocks to new architectures. More specifically, how to rethink the classical algorithms to take into account the dynamics of advanced modern systems. Since a recent past, there have been several papers that propose models for dynamic systems: there is practically a different model for each setting and currently there is no unification of models. Furthermore, models

often suffer of lack of realism. One of the key challenge is to identify which assumptions make sense in new distributed systems. DELYS's objectives are then (1) to identify under which realistic assumptions a given fundamental problem such as mutual exclusion, consensus or leader election can be solved and (2) to design efficient algorithms under these assumptions.

Resource management in heterogeneous systems. The key question is how to manage resources on large and heterogeneous configurations. Managing resources in such systems requires fully decentralized solutions, and to rethink the way various platforms can collaborate and interoperate with each other. In this context, data management is a key component. The fundamental issue we address in ow to efficiently and reliably share information in highly distributed environments.

Adaptation of runtimes. One of the main challenge of the OS community is how to adapt runtime supports to new architectures. With the increasingly widespread use of multicore architectures and virtualised environments, internal runtime protocols need to be revisited. Especially, memory management is crucial in OS and virtualisation technologies have highly impact on it. On one hand, the isolation property of virtualisation has severe side effects on the efficiency of memory allocation since it needs to be constantly balanced between hosted OSs. On the other hand, by hiding the physical machine to OSs, virtualisation prevents them to efficiently place their data in memory on different cores. Our research will thus focus on providing solutions to efficiently share memory between OSs without jeopardizing isolation properties.

DIANA Project-Team

3. Research Program

3.1. Service Transparency

Transparency is to provide network users and application developers with reliable information about the current or predicted quality of their communication services, and about potential leakages of personal information, or of other information related to societal interests of the user as a "connected citizen" (e.g. possible violation of network neutrality, opinion manipulation). Service transparency therefore means to provide information meaningful to users and application developers, such as quality of experience, privacy leakages, or opinion manipulation, etc. rather than network-level metrics such as available bandwidth, loss rate, delay or jitter.

The Internet is built around a best effort routing service that does not provide any guarantee to end users in terms of quality of service (QoS). The simplicity of the Internet routing service is at the root of its huge success. Unfortunately, a simple service means unpredicted quality at the access. Even though a considerable effort is done by operators and content providers to optimise the Internet content delivery chain, mainly by over-provisioning and sophisticated engineering techniques, service degradation is still part of the Internet. The proliferation of wireless and mobile access technologies, and the versatile nature of Internet traffic, make end users quality of experience (QoE) forecast even harder. As a matter of fact, the Internet is missing a dedicated measurement plane that informs the end users on the quality they obtain and in case of substantial service degradation, on the origin of this degradation. Current state of the art activities are devoted to building a distributed measurement infrastructure to perform active, passive and hybrid measurements in the wired Internet. However, the problem is exacerbated with modern terminals such as smartphones or tablets that do not facilitate the task for end users (they even make it harder) as they focus on simplifying the interface and limiting the control on the network, whereas the Internet behind is still the same in terms of the quality it provides. Interestingly, this same observation explains the existing difficulty to detect and prevent privacy leaks. We argue that the lack of transparency for diagnosing QoE and for detecting privacy leaks have the same root causes and can be solved using common primitives. For instance, in both cases, it is important to be able to link data packets to an application. Indeed, as the network can only access data packets, there must be a way to bind these packets to an application (to understand users QoE for this application or to associate a privacy leak to an application). This is however a complex task as the traffic might be obfuscated or encrypted. Our objectives in the research direction are the following:

- Design and develop measurement tools providing transparency, in spite of current complexity
- Deploy those measurement tools at the Internet's edge and make them useful for end users
- Propose measurements plane as an overlay or by exploiting in-network functionalities
- Adapt measurements techniques to network architectural change
- Provide measurements as native functionality in future network architecture

3.2. Open network architecture

We are surrounded by personal content of all types: photos, videos, documents, etc. The volume of such content is increasing at a fast rate, and at the same time, the spread of such content among all our connected devices (mobiles, storage devices, set-top boxes, etc) is also increasing. All this complicates the control of personal content by the user both in terms of access and sharing with other users. The access of the personal content in a seamless way independently of its location is a key challenge for the future of networks. Proprietary solutions exist, but apart from fully depending on one of them, there is no standard plane in the Internet for a seamless access to personal content. Therefore, providing network architectural support to design and develop content access and sharing mechanisms is crucial to allow users control their own data over heterogeneous underlying network or cloud services.

On the other hand, privacy is a growing concern for states, administrations, and companies. Indeed, for instance the French CNIL (entity in charge of citizens privacy in computer systems) puts privacy at the core of its activities by defining rules on any stored and collected private data. Also, companies start to use privacy preserving solutions as a competitive advantage. Therefore, understanding privacy leaks and preventing them is a problem that can already find support. However, all end-users do not *currently* put privacy as their first concern. Indeed, in face of two services with one of higher quality, they usually prefer the highest quality one whatever the privacy implication. This was, for instance, the case concerning the Web search service of Google that is more accurate but less privacy preserving than Bing or Qwant. This is also the case for cloud services such as iCloud or Dropbox that are much more convenient than open source solutions, but very bad in terms of privacy. Therefore, to reach end-users, any privacy preserving solutions must offer a service equivalent to the best existing services.

We consider that it will be highly desirable for Internet users to be able to *easily* move their content from a provider to another and therefore not to depend on a content provider or a social network monopoly. This requires that the network provides built-in architectural support for content networking.

In this research direction, we will define a new *service abstraction layer* (SAL) that could become the new waist of the network architecture with network functionalities below (IP, SDN, cloud) and applications on top. SAL will define different services that are of use to all Internet users for accessing and sharing data (seamless content localisation and retrieval, privacy leakage protection, transparent vertical and horizontal handover, etc.). The biggest challenge here is to cope in the same time with large number of content applications requirements and high underlying networks heterogeneity while still providing efficient applications performance. This requires careful definition of the services primitives and the parameters to be exchanged through the service abstraction layer.

Two concurring factors make the concept behind SAL feasible and relevant today. First, the notion of scalable network virtualization that is a required feature to deploy SAL in real networks today has been discussed recently only. Second, the need for new services abstraction is recent. Indeed, more than fifteen years ago the Internet for the end-users was mostly the Web. Only ten years ago smartphones came into the picture of the Internet boosting the number of applications with new functionalities and risks. Since a few years, many discussions in the network communities took place around the actual complexity of the Internet and the difficulty to develop applications. Many different approaches have been discussed (such as CCN, SDN) that intend to solve only part of the complexity. SAL takes a broader architectural look at the problem and considers solutions such as CCN as mere use cases. Our objectives in this research direction include the following:

- Identify common key networking services required for content access and sharing
- Detect and prevent privacy leaks for content communication
- Enhance software defined networks for large scale heterogeneous environments
- Design and develop open Content Networking architecture
- Define a service abstraction layer as the thin waist for the future content network architecture
- Test and deploy different applications using SAL primitives on heterogeneous network technologies

3.3. Methodology

We follow an experimental approach that can be described in the following techniques:

- Measurements: the aim is to get a better view of a problem in quantifiable terms. Depending on the field of interest, this may involve large scale distributed systems crawling tools; active probing techniques to infer the status and properties of a complex and non controllable system as the Internet; or even crowdsourcing-based deployments for gathering data on real-users environments or behaviours.
- Experimental evaluation: once a new idea has been designed and implemented, it is of course very desirable to assess and quantify how effective it can be, before being able to deploy it on any realistic scale. This is why a wide range of techniques can be considered for getting early, yet as significant as possible, feedback on a given paradigm or implementation. The spectrum for such techniques span from simulations to real deployments in protected and/or controlled environments.

DIONYSOS Project-Team

3. Research Program

3.1. Introduction

The scientific foundations of our work are those of network design and network analysis. Specifically, this concerns the principles of packet switching and in particular of IP networks (protocol design, protocol testing, routing, scheduling techniques), and the mathematical and algorithmic aspects of the associated problems, on which our methods and tools are based.

These foundations are described in the following paragraphs. We begin by a subsection dedicated to Quality of Service (QoS) and Quality of Experience (QoE), since they can be seen as unifying concepts in our activities. Then we briefly describe the specific sub-area of model evaluation and about the particular multidisciplinary domain of network economics.

3.2. Quality of Service and Quality of Experience

Since it is difficult to develop as many communication solutions as possible applications, the scientific and technological communities aim towards providing general *services* allowing to give to each application or user a set of properties nowadays called "Quality of Service" (QoS), a terminology lacking a precise definition. This QoS concept takes different forms according to the type of communication service and the aspects which matter for a given application: for performance it comes through specific metrics (delays, jitter, throughput, etc.), for dependability it also comes through appropriate metrics: reliability, availability, or vulnerability, in the case for instance of WAN (Wide Area Network) topologies, etc.

QoS is at the heart of our research activities: We look for methods to obtain specific "levels" of QoS and for techniques to evaluate the associated metrics. Our ultimate goal is to provide tools (mathematical tools and/or algorithms, under appropriate software "containers" or not) allowing users and/or applications to attain specific levels of QoS, or to improve the provided QoS, if we think of a particular system, with an optimal use of the resources available. Obtaining a good QoS level is a very general objective. It leads to many different areas, depending on the systems, applications and specific goals being considered. Our team works on several of these areas. We also investigate the impact of network QoS on multimedia payloads to reduce the impact of congestion.

Some important aspects of the behavior of modern communication systems have subjective components: the quality of a video stream or an audio signal, *as perceived by the user*, is related to some of the previous mentioned parameters (packet loss, delays, ...) but in an extremely complex way. We are interested in analyzing these types of flows from this user-oriented point of view. We focus on the *user perceived quality*, in short, PQ, the main component of what is nowadays called Quality of Experience (in short, QoE), to underline the fact that, in this case, we want to center the analysis on the user. In this context, we have a global project called PSQA, which stands for Pseudo-Subjective Quality Assessment, and which refers to a technology we have developed allowing to automatically measure this PQ.

Another special case to which we devote research efforts in the team is the analysis of qualitative properties related to interoperability assessment. This refers to the act of determining if end-to-end functionality between at least two communicating systems is as required by the base standards for those systems. Conformance is the act of determining to what extent a single component conforms to the individual requirements of the standard it is based on. Our purpose is to provide such a formal framework (methods, algorithms and tools) for interoperability assessment, in order to help in obtaining efficient interoperability test suites for new generation networks, mainly around IPv6-related protocols. The interoperability test suites generation is based on specifications (standards and/or RFCs) of network components and protocols to be tested.

3.3. Stochastic modeling

The scientific foundations of our modeling activities are composed of stochastic processes theory and, in particular, Markov processes, queuing theory, stochastic graphs theory, etc. The objectives are either to develop numerical solutions, or analytical ones, or possibly discrete event simulation or Monte Carlo (and Quasi-Monte Carlo) techniques. We are always interested in model evaluation techniques for dependability and performability analysis, both in static (network reliability) and dynamic contexts (depending on the fact that time plays an explicit role in the analysis or not). We look at systems from the classical so-called *call level*, leading to standard models (for instance, queues or networks of queues) and also at the *burst level*, leading to *fluid models*.

In recent years, our work on the design of the topologies of WANs led us to explore optimization techniques, in particular in the case of very large optimization problems, usually formulated in terms of graphs. The associated methods we are interested in are composed of simulated annealing, genetic algorithms, TABU search, etc. For the time being, we have obtained our best results with GRASP techniques.

Network pricing is a good example of a multi-disciplinary research activity half-way between applied mathematics, economy and networking, centered on stochastic modeling issues. Indeed, the Internet is facing a tremendous increase of its traffic volume. As a consequence, real users complain that large data transfers take too long, without any possibility to improve this by themselves (by paying more, for instance). A possible solution to cope with congestion is to increase the link capacities; however, many authors consider that this is not a viable solution as the network must respond to an increasing demand (and experience has shown that demand of bandwidth has always been ahead of supply), especially now that the Internet is becoming a commercial network. Furthermore, incentives for a fair utilization between customers are not included in the current Internet. For these reasons, it has been suggested that the current flat-rate fees, where customers pay a subscription and obtain an unlimited usage, should be replaced by usage-based fees. Besides, the future Internet will carry heterogeneous flows such as video, voice, email, web, file transfers and remote login among others. Each of these applications requires a different level of QoS: for example, video needs very small delays and packet losses, voice requires small delays but can afford some packet losses, email can afford delay (within a given bound) while file transfer needs a good average throughput and remote login requires small round-trip times. Some pricing incentives should exist so that each user does not always choose the best QoS for her application and so that the final result is a fair utilization of the bandwidth. On the other hand, we need to be aware of the trade-off between engineering efficiency and economic efficiency; for example, traffic measurements can help in improving the management of the network but is a costly option. These are some of the various aspects often present in the pricing problems we address in our work. More recently, we have switched to the more general field of network economics, dealing with the economic behavior of users, service providers and content providers, as well as their relations.

DIVERSE Project-Team

3. Research Program

3.1. Scientific background

3.1.1. Model-Driven Engineering

Model-Driven Engineering (MDE) aims at reducing the accidental complexity associated with developing complex software-intensive systems (e.g., use of abstractions of the problem space rather than abstractions of the solution space) [120]. It provides DIVERSE with solid foundations to specify, analyze and reason about the different forms of diversity that occur through the development lifecycle. A primary source of accidental complexity is the wide gap between the concepts used by domain experts and the low-level abstractions provided by general-purpose programming languages [91]. MDE approaches address this problem through modeling techniques that support separation of concerns and automated generation of major system artifacts from models (*e.g.,* test cases, implementations, deployment and configuration scripts). In MDE, a model describes an aspect of a system and is typically created or derived for specific development purposes [73]. Separation of concerns is supported through the use of different modeling languages, each providing constructs based on abstractions that are specific to an aspect of a system. MDE technologies also provide support for manipulating models, for example, support for querying, slicing, transforming, merging, and analyzing (including executing) models. Modeling languages are thus at the core of MDE, which participates in the development of a sound *Software Language Engineering* [0], including a unified typing theory that integrate models as first class entities [123].

Incorporating domain-specific concepts and high-quality development experience into MDE technologies can significantly improve developer productivity and system quality. Since the late nineties, this realization has led to work on MDE language workbenches that support the development of domain-specific modeling languages (DSMLs) and associated tools (*e.g.,* model editors and code generators). A DSML provides a bridge between the field in which domain experts work and the implementation (programming) field. Domains in which DSMLs have been developed and used include, among others, automotive, avionics, and the emerging cyber-physical systems. A study performed by Hutchinson et al. [97] indicates that DSMLs can pave the way for wider industrial adoption of MDE.

More recently, the emergence of new classes of systems that are complex and operate in heterogeneous and rapidly changing environments raises new challenges for the software engineering community. These systems must be adaptable, flexible, reconfigurable and, increasingly, self-managing. Such characteristics make systems more prone to failure when running and thus development and study of appropriate mechanisms for continuous design and runtime validation and monitoring are needed. In the MDE community, research is focused primarily on using models at design, implementation, and deployment stages of development. This work has been highly productive, with several techniques now entering a commercialization phase. As software systems are becoming more and more dynamic, the use of model-driven techniques for validating and monitoring runtime behavior is extremely promising [105].

3.1.2. Variability modeling

While the basic vision underlying *Software Product Lines* (SPL) can probably be traced back to David Parnas' seminal article [113] on the Design and Development of Program Families, it is only quite recently that SPLs are emerging as a paradigm shift towards modeling and developing software system families rather than individual systems [111]. SPL engineering embraces the ideas of mass customization and software reuse. It focuses on the means of efficiently producing and maintaining multiple related software products, exploiting what they have in common and managing what varies among them.

[0] See http://planet-sl.org

Several definitions of the *software product line* concept can be found in the research literature. Clements *et al.* define it as *a set of software-intensive systems sharing a common, managed set of features that satisfy the specific needs of a particular market segment or mission and are developed from a common set of core assets in a prescribed way* [110]. Bosch provides a different definition [79]: *A SPL consists of a product line architecture and a set of reusable components designed for incorporation into the product line architecture. In addition, the PL consists of the software products developed using the mentioned reusable assets*. In spite of the similarities, these definitions provide different perspectives of the concept: *market-driven*, as seen by Clements *et al.*, and *technology-oriented* for Bosch.

SPL engineering is a process focusing on capturing the *commonalities* (assumptions true for each family member) and *variability* (assumptions about how individual family members differ) between several software products [85]. Instead of describing a single software system, a SPL model describes a set of products in the same domain. This is accomplished by distinguishing between elements common to all SPL members, and those that may vary from one product to another. Reuse of core assets, which form the basis of the product line, is key to productivity and quality gains. These core assets extend beyond simple code reuse and may include the architecture, software components, domain models, requirements statements, documentation, test plans or test cases.

The SPL engineering process consists of two major steps:

1. **Domain Engineering**, or *development for reuse*, focuses on core assets development.
2. **Application Engineering**, or *development with reuse*, addresses the development of the final products using core assets and following customer requirements.

Central to both processes is the management of **variability** across the product line [93]. In common language use, the term *variability* refers to *the ability or the tendency to change*. Variability management is thus seen as the key feature that distinguishes SPL engineering from other software development approaches [80]. Variability management is thus growingly seen as the cornerstone of SPL development, covering the entire development life cycle, from requirements elicitation [125] to product derivation [130] to product testing [109], [108].

Halmans *et al.* [93] distinguish between *essential* and *technical* variability, especially at requirements level. Essential variability corresponds to the customer's viewpoint, defining what to implement, while technical variability relates to product family engineering, defining how to implement it. A classification based on the dimensions of variability is proposed by Pohl *et al.* [115]: beyond **variability in time** (existence of different versions of an artifact that are valid at different times) and **variability in space** (existence of an artifact in different shapes at the same time) Pohl *et al.* claim that variability is important to different stakeholders and thus has different levels of visibility: **external variability** is visible to the customers while **internal variability**, that of domain artifacts, is hidden from them. Other classification proposals come from Meekel *et al.* [103] (feature, hardware platform, performances and attributes variability) or Bass *et al.* [71] who discusses about variability at the architectural level.

Central to the modeling of variability is the notion of *feature*, originally defined by Kang *et al.* as: *a prominent or distinctive user-visible aspect, quality or characteristic of a software system or systems* [99]. Based on this notion of *feature*, they proposed to use a *feature model* to model the variability in a SPL. A feature model consists of a *feature diagram* and other associated information: *constraints* and *dependency rules*. Feature diagrams provide a *graphical tree-like notation depicting the hierarchical organization of high level product functionalities* represented as features. The root of the tree refers to the complete system and is progressively decomposed into more refined features (tree nodes). Relations between nodes (features) are materialized by *decomposition edges* and *textual constraints*. Variability can be expressed in several ways. Presence or absence of a feature from a product is modeled using *mandatory* or *optional features*. Features are graphically represented as rectangles while some graphical elements (e.g., unfilled circle) are used to describe the variability (e.g., a feature may be optional).

Features can be organized into *feature groups*. Boolean operators *exclusive alternative (XOR)*, *inclusive alternative (OR)* or *inclusive (AND)* are used to select one, several or all the features from a feature group.

Dependencies between features can be modeled using *textual constraints*: *requires* (presence of a feature requires the presence of another), *mutex* (presence of a feature automatically excludes another). Feature attributes can be also used for modeling quantitative (e.g., numerical) information. Constraints over attributes and features can be specified as well.

Modeling variability allows an organization to capture and select which version of which variant of any particular aspect is wanted in the system [80]. To implement it cheaply, quickly and safely, redoing by hand the tedious weaving of every aspect is not an option: some form of automation is needed to leverage the modeling of variability [75], [87]. Model Driven Engineering (MDE) makes it possible to automate this weaving process [98]. This requires that models are no longer informal, and that the weaving process is itself described as a program (which is as a matter of facts an executable meta-model [106]) manipulating these models to produce for instance a detailed design that can ultimately be transformed to code, or to test suites [114], or other software artifacts.

3.1.3. Component-based software development

Component-based software development [124] aims at providing reliable software architectures with a low cost of design. Components are now used routinely in many domains of software system designs: distributed systems, user interaction, product lines, embedded systems, etc. With respect to more traditional software artifacts (e.g., object oriented architectures), modern component models have the following distinctive features [86]: description of requirements on services required from the other components; indirect connections between components thanks to ports and connectors constructs [101]; hierarchical definition of components (assemblies of components can define new component types); connectors supporting various communication semantics [83]; quantitative properties on the services [78].

In recent years component-based architectures have evolved from static designs to dynamic, adaptive designs (e.g., SOFA [83], Palladio [76], Frascati [107]). Processes for building a system using a statically designed architecture are made of the following sequential lifecycle stages: requirements, modeling, implementation, packaging, deployment, system launch, system execution, system shutdown and system removal. If for any reason after design time architectural changes are needed after system launch (e.g., because requirements changed, or the implementation platform has evolved, etc) then the design process must be reexecuted from scratch (unless the changes are limited to parameter adjustment in the components deployed).

Dynamic designs allow for *on the fly* redesign of a component based system. A process for dynamic adaptation is able to reapply the design phases while the system is up and running, without stopping it (this is different from a stop/redeploy/start process). Dynamic adaptation process supports *chosen adaptation*, when changes are planned and realized to maintain a good fit between the needs that the system must support and the way it supports them [100]. Dynamic component-based designs rely on a component meta-model that supports complex life cycles for components, connectors, service specification, etc. Advanced dynamic designs can also take platform changes into account at runtime, without human intervention, by adapting themselves [84], [127]. Platform changes and more generally environmental changes trigger *imposed adaptation*, when the system can no longer use its design to provide the services it must support. In order to support an eternal system [77], dynamic component based systems must separate architectural design and platform compatibility. This requires support for heterogeneity, since platform evolution can be partial.

The Models@runtime paradigm denotes a model-driven approach aiming at taming the complexity of dynamic software systems. It basically pushes the idea of reflection one step further by considering the reflection layer as a real model "something simpler, safer or cheaper than reality to avoid the complexity, danger and irreversibility of reality [118]". In practice, component-based (and/or service-based) platforms offer reflection APIs that make it possible to introspect the system (to determine which components and bindings are currently in place in the system) and dynamic adaptation (by applying CRUD operations on these components and bindings). While some of these platforms offer rollback mechanisms to recover after an erroneous adaptation, the idea of Models@runtime is to prevent the system from actually enacting an erroneous adaptation. In other words, the "model at run-time" is a reflection model that can be uncoupled (for reasoning, validation, simulation purposes) and automatically resynchronized.

Heterogeneity is a key challenge for modern component based system. Until recently, component based techniques were designed to address a specific domain, such as embedded software for command and control, or distributed Web based service oriented architectures. The emergence of the Internet of Things paradigm calls for a unified approach in component based design techniques. By implementing an efficient separation of concern between platform independent architecture management and platform dependent implementations, *Models@runtime* is now established as a key technique to support dynamic component based designs. It provides DIVERSE with an essential foundation to explore an adaptation envelop at run-time.

Search Based Software Engineering [95] has been applied to various software engineering problems in order to support software developers in their daily work. The goal is to automatically explore a set of alternatives and assess their relevance with respect to the considered problem. These techniques have been applied to craft software architecture exhibiting high quality of services properties [92]. Multi Objectives Search based techniques [89] deal with optimization problem containing several (possibly conflicting) dimensions to optimize. These techniques provide DIVERSE with the scientific foundations for reasoning and efficiently exploring an envelope of software configurations at run-time.

3.1.4. Validation and verification

Validation and verification (V&V) theories and techniques provide the means to assess the validity of a software system with respect to a specific correctness envelop. As such, they form an essential element of DIVERSE's scientific background. In particular, we focus on model-based V&V in order to leverage the different models that specify the envelop at different moments of the software development lifecycle.

Model-based testing consists in analyzing a formal model of a system (*e.g.*, activity diagrams, which capture high-level requirements about the system, statecharts, which capture the expected behavior of a software module, or a feature model, which describes all possible variants of the system) in order to generate test cases that will be executed against the system. Model-based testing [126] mainly relies on model analysis, constraint solving [88] and search-based reasoning [102]. DIVERSE leverages in particular the applications of model-based testing in the context of highly-configurable systems and [128] interactive systems [104] as well as recent advances based on diversity for test cases selection [96].

Nowadays, it is possible to simulate various kinds of models. Existing tools range from industrial tools such as Simulink, Rhapsody or Telelogic to academic approaches like Omega [112], or Xholon [0]. All these simulation environments operate on homogeneous environment models. However, to handle diversity in software systems, we also leverage recent advances in heterogeneous simulation. Ptolemy [82] proposes a common abstract syntax, which represents the description of the model structure. These elements can be decorated using different directors that reflect the application of a specific model of computation on the model element. Metropolis [72] provides modeling elements amenable to semantically equivalent mathematical models. Metropolis offers a precise semantics flexible enough to support different models of computation. ModHel'X [94] studies the composition of multi-paradigm models relying on different models of computation.

Model-based testing and simulation are complemented by runtime fault-tolerance through the automatic generation of software variants that can run in parallel, to tackle the open nature of software-intensive systems. The foundations in this case are the seminal work about N-version programming [70], recovery blocks [116] and code randomization [74], which demonstrated the central role of diversity in software to ensure runtime resilience of complex systems. Such techniques rely on truly diverse software solutions in order to provide systems with the ability to react to events, which could not be predicted at design time and checked through testing or simulation.

3.1.5. Empirical software engineering

The rigorous, scientific evaluation of DIVERSE's contributions is an essential aspect of our research methodology. In addition to theoretical validation through formal analysis or complexity estimation, we also aim at applying state-of-the-art methodologies and principles of empirical software engineering. This approach encompasses a set of techniques for the sound validation contributions in the field of software engineering,

[0] http://www.primordion.com/Xholon/

ranging from statistically sound comparisons of techniques and large-scale data analysis to interviews and systematic literature reviews [121], [119]. Such methods have been used for example to understand the impact of new software development paradigms [81]. Experimental design and statistical tests represent another major aspect of empirical software engineering. Addressing large-scale software engineering problems often requires the application of heuristics, and it is important to understand their effects through sound statistical analyses [69].

3.2. Research axis

Figure 1 illustrates the four dimensions of software diversity, which form the core research axis of DIVERSE: the **diversity of languages** used by the stakeholders involved in the construction of these systems; the **diversity of features** required by the different customers; the **diversity of runtime environments** in which software has to run and adapt; the **diversity of implementations** that are necessary for resilience through redundancy. These four axes share and leverage the scientific and technological results developed in the area of model-driven engineering in the last decade. This means that all our research activities are founded on sound abstractions to reason about specific aspects of software systems, compose different perspectives and automatically generate parts of the system.

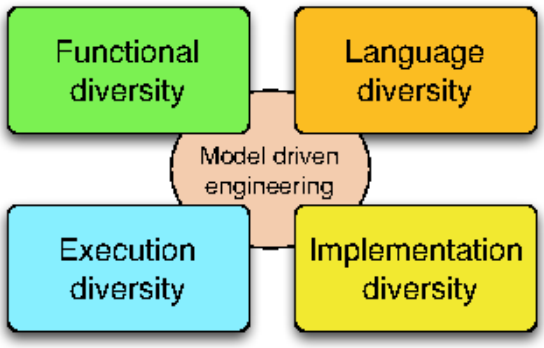


Figure 1. The four research axes of DIVERSE, which rely on a MDE scientific background

3.2.1. Software Language Engineering

The engineering of systems involves many different stakeholders, each with their own domain of expertise. Hence more and more organizations are adopting Domain Specific Modeling Languages (DSMLs) to allow domain experts to express solutions directly in terms of relevant domain concepts [120], [91]. This new trend raises new challenges about designing DSMLs, evolving a set of DSMLs and coordinating the use of multiple DSLs for both DSL designers and DSL users.

3.2.1.1. Challenges

Reusability of software artifacts is a central notion that has been thoroughly studied and used by both academics and industrials since the early days of software construction. Essentially, designing reusable artifacts allows the construction of large systems from smaller parts that have been separately developed and validated, thus reducing the development costs by capitalizing on previous engineering efforts. However, it is still hardly possible for language designers to design typical language artifacts (e.g. language constructs, grammars, editors or compilers) in a reusable way. The current state of the practice usually prevents the reusability of language artifacts from one language to another, consequently hindering the emergence of real engineering techniques around software languages. Conversely, concepts and mechanisms that enable artifacts reusability abound in the software engineering community.

Variability in modeling languages occur in the definition of the abstract and concrete syntax as well as in the specification of the language's semantics. The major challenges met when addressing the need for variability are: (i) to set principles for modeling language units that support the modular specification of a modeling language; and (ii) to design mechanisms to assemble these units into a complete language, according to the set of authorized variation points for the modeling language family.

A new generation of complex software-intensive systems (for example smart health support, smart grid, building energy management, and intelligent transportation systems) gives new opportunities for leveraging modeling languages. The development of these systems requires expertise in diverse domains. Consequently, different types of stakeholders (e.g., scientists, engineers and end-users) must work in a coordinated manner on various aspects of the system across multiple development phases. DSMLs can be used to support the work of domain experts who focus on a specific system aspect, but they can also provide the means for coordinating work across teams specializing in different aspects and across development phases. The support and integration of DSMLs leads to what we call **the globalization of modeling languages**, *i.e.* the use of multiple languages for the coordinated development of diverse aspects of a system. One can make an analogy with world globalization in which relationships are established between sovereign countries to regulate interactions (e.g., travel and commerce related interactions) while preserving each country's independent existence.

3.2.1.2. Scientific objectives

We address reuse and variability challenges through the investigation of the time-honored concepts of substitutability, inheritance and components, evaluate their relevance for language designers and provide tools and methods for their inclusion in software language engineering. We will develop novel techniques for the modular construction of language extensions with support to model syntactical variability. From the semantics perspective, we investigate extension mechanisms for the specification of variability in operational semantics, focusing on static introduction and heterogeneous models of computation. The definition of variation points for the three aspects of the language definition provides the foundations for the novel concept Language Unit (LU) as well as suitable mechanisms to compose such units.

We explore the necessary breakthrough in software languages to support modeling and simulation of heterogeneous and open systems. This work relies on the specification of executable domain specific modeling languages (DSMLs) to formalize the various concerns of a software-intensive system, and of models of computation (MoCs) to explicitly model the concurrency, time and communication of such DSMLs. We develop a framework that integrates the necessary foundations and facilities for designing and implementing executable and concurrent domain-specific modeling languages. This framework also provides unique features to specify composition operators between (possibly heterogeneous) DSMLs. Such specifications are amenable to support the edition, execution, graphical animation and analysis of heterogeneous models. The objective is to provide both a significant improvement to MoCs and DSMLs design and implementation and to the simulation based validation and verification of complex systems.

We see an opportunity for the automatic diversification of programs' computation semantics, for example through the diversification of compilers or virtual machines. The main impact of this artificial diversity is to provide flexible computation and thus ease adaptation to different execution conditions. A combination of static and dynamic analysis could support the identification of what we call *plastic computation zones* in the code. We identify different categories of such zones: (i) areas in the code in which the order of computation can vary (e.g., the order in which a block of sequential statements is executed); (ii) areas that can be removed, keeping the essential functionality [122] (e.g., skip some loop iterations); (iii) areas that can replaced by alternative code (e.g., replace a try-catch by a return statement). Once we know which zones in the code can be randomized, it is necessary to modify the model of computation to leverage the computation plasticity. This consists in introducing variation points in the interpreter to reflect the diversity of models of computation. Then, the choice of a given variation is performed randomly at run time.

3.2.2. Variability Modeling and Engineering

The systematic modeling of variability in software systems has emerged as an effective approach to document and reason about software evolution and heterogeneity (*cf.* Section 3.1.2). Variability modeling characterizes

an "envelope" of possible software variations. The industrial use of variability models and their relation to software artifact models require a complete engineering framework, including composition, decomposition, analysis, configuration and artifact derivation, refactoring, re-engineering, extraction, and testing. This framework can be used both to tame imposed diversity and to manage chosen diversity.

3.2.2.1. Challenges

A fundamental problem is that the **number of variants** can be exponential in the number of options (features). Already with 300 boolean configuration options, approximately 10^{90} configurations exist – more than the estimated count of atoms in the universe. Domains like automotive or operating systems have to manage more than 10000 options (e.g., Linux). Practitioners face the challenge of developing billions of variants. It is easy to forget a necessary constraint, leading to the synthesis of unsafe variants, or to under-approximate the capabilities of the software platform. Scalable modelling techniques are therefore crucial to specify and reason about a very large set of variants.

Model-driven development supports two approaches to deal with the increasing number of concerns in complex systems: multi-view modeling, *i.e.* when modeling each concern separately, and variability modeling. However, there is little support to combine both approaches consistently. Techniques to integrate both approaches will enable the construction of a consistent set of views and variation points in each view.

The design, construction and maintenance of software families have a major impact on **software testing**. Among the existing challenges, we can cite: the selection of test cases for a specific variant; the evolution of test suites with integration of new variants; the combinatorial explosion of the number of software configurations to be tested. Novel model-based techniques for test generation and test management in a software product line context are needed to overcome state-of-the-art limits we already observed in some projects.

3.2.2.2. Scientific objectives

We aim at developing scalable reasoning techniques to **automatically analyze** variability models and their interactions with other views on the software intensive system (requirements, architecture, design, code). These techniques provide two major advancements in the state of the art: (1) an extension of the semantics of variability models in order to enable the definition of attributes (*e.g.*, cost, quality of service, effort) on features and to include these attributes in the reasoning; (2) an assessment of the consistent specification of variability models with respect to system views (since variability is orthogonal to system modeling, it is currently possible to specify the different models in ways that are semantically meaningless). The former aspect of analysis is tackled through constraint solving and finite-domain constraint programming, while the latter aspect is investigated through automatic search-based and learning-based techniques for the exploration of the space of interaction between variability and view models.

We aim at developing procedures to **reverse engineer** dependencies and features' sets from existing software artefacts – be it source code, configuration files, spreadsheets (e.g., product comparison matrices) or requirements. We expect to scale up (e.g., for extracting a very large number of variation points) and guarantee some properties (e.g., soundness of configuration semantics, understandability of ontological semantics). For instance, when building complex software-intensive systems, textual requirements are captured in very large quantities of documents. In this context, adequate models to formalize the organization of requirements documents and automated techniques to support impact analysis (in case of changes in the requirements) have to be developed.

3.2.3. Heterogeneous and dynamic software architectures

Flexible yet dependable systems have to cope with heterogeneous hardware execution platforms ranging from smart sensors to huge computation infrastructures and data centers. Evolution possibilities range from a mere change in the system configuration to a major architectural redesign, for instance to support addition of new features or a change in the platform architecture (e.g., new hardware is made available, a running system switches to low bandwidth wireless communication, a computation node battery is running low, etc). In this context, we need to devise formalisms to reason about the impact of an evolution and about the transition from one configuration to another. It must be noted that this axis focuses on the use of models to drive the evolution

from design time to runtime. Models will be used to (i) systematically define predictable configurations and variation points through which the system will evolve; (ii) develop behaviors necessary to handle unforeseen evolution cases.

3.2.3.1. Challenges

The main challenge is to provide new homogeneous architectural modelling languages and efficient techniques that enable continuous software reconfiguration to react to changes. This work handles the challenges of handling the diversity of runtime infrastructures and managing the cooperation between different stakeholders. More specifically, the research developed in this axis targets the following dimensions of software diversity.

Platform architectural heterogeneity induces a first dimension of imposed diversity (type diversity). Platform reconfiguration driven by changing resources define another dimension of diversity (deployment diversity). To deal with these imposed diversity problems, we will rely on model based runtime support for adaptation, in the spirit of the dynamic distributed component framework developed by the Triskell team. Since the runtime environment composed of distributed, resource constrained hardware nodes cannot afford the overhead of traditional runtime adaptation techniques, we investigate the design of novel solutions relying on Models@runtime and on specialized tiny virtual machines to offer resource provisioning and dynamic reconfiguration.

Diversity can also be an asset to optimize software architecture. Architecture models must integrate multiple concerns in order to properly manage the deployment of software components over a physical platform. However, these concerns can contradict each other (*e.g.*, accuracy and energy). In this context, we investigate automatic solutions to explore the set of possible architecture models and to establish valid trade-offs between all concerns in case of changes.

3.2.3.2. Scientific objectives

Automatic synthesis of optimal software architectures. Implementing a service over a distributed platform (*e.g.*, a pervasive system or a cloud platform) consists in deploying multiple software components over distributed computation nodes. We aim at designing search-based solutions to (i) assist the software architect in establishing a good initial architecture (that balances between different factors such as cost of the nodes, latency, fault tolerance) and to automatically update the architecture when the environment or the system itself change. The choice of search-based techniques is motivated by the very large number of possible software deployment architectures that can be investigated and that all provide different trade-offs between qualitative factors. Another essential aspect that is supported by multi-objective search is to explore different architectural solutions that are not necessarily comparable. This is important when the qualitative factors are orthogonal to each other, such as security and usability for example.

Flexible software architecture for testing and data management. As the number of platforms on which software runs increases and different software versions coexist, the demand for testing environments also increases. For example, the number of testing environments to test a software patch or upgrade is the product of the number of execution environments the software supports and the number of coexisting versions of the software. Based on our first experiment on the synthesis of cloud environment using architectural models, our objective is to define a set of domain specific languages to catch the requirement and to design cloud environments for testing and data management of future internet systems from data centers to things. These languages will be interpreted to support dynamic synthesis and reconfiguration of a testing environment.

Runtime support for heterogeneous environments. Execution environments must provide a way to account or reserve resources for applications. However, current execution environments such as the Java Virtual Machine do not clearly define a notion of application: each framework has its own definition. For example, in OSGi, an application is a component, in JEE, an application is most of the time associated to a class loader, in the Multi-Tasking Virtual machine, an application is a process. The challenge consists in defining an execution environment that provides direct control over resources (CPU, Memory, Network I/O) independently from the definition of an application. We propose to define abstract resource containers to account and reserve resources on a distributed network of heterogeneous devices.

3.2.4. Diverse implementations for resilience

Open software-intensive systems have to evolve over their lifetime in response to changes in their environment. Yet, most verification techniques assume a closed environment or the ability to predict all changes. Dynamic changes and evolution cases thus represent a major challenge for these techniques that aim at assessing the correctness and robustness of the system. On the one hand, DIVERSE will adapt V&V techniques to handle diversity imposed by the requirements and the execution environment, on the other hand we leverage diversity to increase the robustness of software in face of unforeseen situations. More specifically, we address the following V&V challenges.

3.2.4.1. Challenges

One major challenge to build flexible and open yet dependable systems is that current software engineering techniques require architects to foresee all possible situations the system will have to face. However, openness and flexibility also mean unpredictability: unpredictable bugs, attacks, environmental evolution, etc. Current fault-tolerance [116] and security [90] techniques provide software systems with the capacity of detecting accidental and deliberate faults. However, existing solutions assume that the set of bugs or vulnerabilities in a system does not evolve. This assumption does not hold for open systems, thus it is essential to revisit fault-tolerance and security solutions to account for diverse and unpredictable faults.

Diversity is known to be a major asset for the robustness of large, open, and complex systems (*e.g.*, economical or ecological systems). Following this observation, the software engineering literature provides a rich set of work that rely on implementation diversity in software systems in order to improve robustness to attacks or to changes in quality of service. These works range from N-version programming to obfuscation of data structures or control flow, to randomization of instruction sets. An essential and active challenge is to support the automatic synthesis and evolution of software diversity in open software-intensive systems. There is an opportunity to further enhance these techniques in order to cope with a wider diversity of faults, by multiplying the levels of diversity in the different software layers that are found in software-intensive systems (system, libraries, frameworks, application). This increased diversity must be based on artificial program transformations and code synthesis, which increase the chances of exploring novel solutions, better fitted at one point in time. The biological analogy also indicates that diversity should emerge as a side-effect of evolution, to prevent over-specialization towards one kind of diversity.

3.2.4.2. Scientific objectives

The main objective is to address one of the main limitations of N-version programming for fault-tolerant systems: the manual production and management of software diversity. Through automated injection of artificial diversity we aim at systematically increasing failure diversity and thus increasing the chances of early error detection at run-time. A fundamental assumption for this work is that software-intensive systems can be “good enough” [117], [129].

Proactive program diversification. We aim at establishing novel principles and techniques that favor the emergence of multiple forms of software diversity in software-intensive systems, in conjunction with the software adaptation mechanisms that leverage this diversity. The main expected outcome is a set of meta-design principles that maintain diversity in systems and the experimental demonstration of the effects of software diversity. Higher levels of diversity in the system provide a pool of software solutions that can eventually be used to adapt to situations unforeseen at design time (bugs, crash, attacks, etc.). Principles of automated software diversification rely on the automated synthesis of variants in a software product line, as well as finer-grained program synthesis combining unsound transformations and genetic programming to explore the space of mutational robustness.

Multi-tier software diversification. We name multi-tier diversification the fact of diversifying several application software components simultaneously. The novelty of our proposal, with respect to the software diversity state of the art, is to diversify the application-level code (for example, diversify the business logic of the application), focusing on the technical layers found in web applications. The diversification of application software code is expected to provide a diversity of failures and vulnerabilities in web server deployment. Web server deployment usually adopts a form of the Reactor architecture pattern, for scalability purposes:

multiple copies of the server software stack, called request handlers, are deployed behind a load balancer. This architecture is very favorable for diversification, since by using the multiplicity of request handlers running in a web server we can simultaneously deploy multiple combinations of diverse software components. Then, if one handler is hacked or crashes the others should still be able to process client requests.

DYOGENE Project-Team

3. Research Program

3.1. Initial research axes

The following research axes have been defined in 2013 when the project-team was created.

- Algorithms for network performance analysis, led by A. Bouillard and A. Busic.
- Stochastic geometry and information theory for wireless network, led by F. Baccelli and B. Blaszczyszyn.
- The cavity method for network algorithms, led by M. Lelarge.

Our scientific interests keep evolving. Research areas which received the most of our attention in 2019 are summarized in the following sections.

3.2. Distributed network control and smart-grids

Theory and algorithms for distributed control of networks with applications to the stabilization of power grids subject to high volatility of renewable energy production are being developed by A. Busic in collaboration with Sean Meyn [Prof. at University of Florida and Inria International Chair].

3.3. Mathematics of wireless cellular networks

A comprehensive approach involving information theory, queueing and stochastic geometry to model and analyze the performance of large cellular networks, validated and implemented by Orange is being led by B. Blaszczyszyn in collaboration with F. Baccelli and M. K. Karray [Orange Labs]. A new collaboration between the Standardization and Research Lab at Nokia Bell Labs and ERC NEMO led by F. Baccelli has been started in 2019.

3.4. High-dimensional statistical inference and distributed learning

We computed information theoretic bounds for unsupervised and semi-supervised learning and proved complexity bounds for distributed optimization of convex functions using a network of computing units.

3.5. Stochastic Geometry

In collaboration with Mir-Omid Haji-Mirsadeghi [Sharif University, Tehran, Iran] and Ali Khezeli [School of Mathematical Sciences, Tehran, Iran] F. Baccelli develops a theory of unimodular random metric spaces.

The distortion properties of unconstrained one-bit compression were analyzed by F. Baccelli in collaboration with E. O'Reilly [Caltech] using high dimensional hyperplane tessellations.

In collaboration with D. Yogeshwaran [Indian Statistical Institute, Bangalore] and J. E. Yukich [Lehigh University] B. Błaszczyszyn develops a limit theory (Laws of Large Numbers and Central Limit Theorems) for functionals of spatially correlated point processes.

EASE Project-Team

3. Research Program

3.1. Collecting pertinent information

In our model, applications adapt their behavior (for instance, the level of automation) to the quality of their perception of the environment. This is important to alleviate the development constraint we usually have on automated systems. We "just" have to be sure a given process will always operate at the right automation level given the precision, the completeness or the confidence it has on its own perception. For instance, a car passing through crossing would choose its speed depending on the confidence it has gained during perception data gathering. When it has not enough information or when it could not trust it, it should reduce the automation level, therefore the speed, to only rely on its own sensors. Such adaptation capability shift requirements from the design and deployment (availability, robustness, accuracy, etc.) to the **assessment of the environment perception** we aim to facilitate in this first research axis.

Data characterization. The quality (freshness, accuracy, confidence, reliability, confidentiality, etc.) of the data are of crucial importance to assess the quality of the perception and therefore to ensure proper behavior. The way data is produced, consolidated, and aggregated while flowing to the consumer has an impact on its quality. Moreover part of these quality attributes requires to gather information at several communication layers from various entities. For this purpose, we want to design **lightweight cross-layer interactions** to collect relevant data. As a "frugality" principle should guide our approach, it is not appropriate to build all attributes we can imagine. It is therefore necessary to identify attributes relevant to the application and to have mechanisms to activate/deactivate at run-time the process to collect them.

Data fusion. Raw data should be directly used only to determine low-level abstraction. Further help in abstracting from low-level details can be provided by **data fusion** mechanisms. A good (re)construction of a meaningful information for the application reduces the complexity of the pervasive applications and helps the developers to concentrate on the application logic rather on the management of raw data. Moreover, the reactivity required in pervasive systems and the aggregation of large amounts of data (and its processing) are antagonists. We study **software services that can be deployed closer to the edge of the network**. The exploration of data fusion technics will be guided by different criteria: relevance of abstractions produced for pervasive applications, anonymization of exploited raw data, processing time, etc.

Assessing the correctness of the behavior. To ease the design of new applications and to align the development of new products with the ever faster standard developments, continuous integration could be used in parallel with continuous conformance and interoperability testing. We already participate in the design of new shared platforms that aims at facilitating this providing remote testing tools. Unfortunately, it is not possible to be sure that all potential peers in the surrounding have a conform behavior. Moreover, upon failure or security breach, a piece of equipment could stop to operate properly and lead to global mis-behavior. We want to propose conceptual tools for **testing at runtime devices in the environment**. The result of such conformance or interoperability tests could be stored safely in the environment by authoritative testing entity. Then application could interact with the device with a higher confidence. The confidence level of a device could be part of the quality attribute of the information it contributed to generate. The same set of tools could be used to identify misbehaving device for maintenance purpose or to trigger further testing.

3.2. Building relevant abstraction for new interactions

The pervasive applications are often designed in an ad hoc manner depending on the targeted application area. Ressources (sensors / actuators, connected objets etc.) are often used in silos which complexify the implementation of rich pervasive computing scenarios. In the second research axis, we want to get away from technical aspects identifying **common and reusable system mechanisms** that could be used in various applications.

Tagging the environment. Information relative to environment could be stored by the application itself, but it could be complex to manage for mobile application since it could cross a large number of places with various features. Moreover the developer has to build its own representation of information especially when he wants to share information with other instances of the same application or with other applications. A promising approach is to store and to maintain this information associated to an object or to a place, in the environment itself. The infrastructure should provide services to application developers: add/retrieve information in the environment, share information and control who can access it, add computed properties to object for further usage. We want to study an **extensible model to describe and augment the environment**. Beyond a simple distributed storage, we have in mind a new kind of interaction between pervasive applications and changing environment and between applications themselves.

Taking advantages of the spatial relationships. To understand the world they have to interact with, pervasive applications often have to (re)built a model of it from the exchange they have with others or from their own observations. A part of the programmer's task consists in building a model of the spatial layout of the objects in the surrounding. The term *layout* can be understood in several ways: the co-location of multiple objects in the same vicinity, the physical arrangement of two objects relative to each other, or even the crossing of an object of a physical area to another, etc. Determining remotely these spatial properties (see figure 1 -a) is difficult without exchanging a lot of information. Properties related to the spatial layout are far easier to characterize locally. They could be abstracted from interaction pattern without any complex virtual representation of the environment (see figure 1 -b). We want to be able to rely on this type of spatial layout in a pervasive environment. In the prior years, the members of EASE already worked on **models for processing object interactions** in the physical world to automatically trigger processing. This was the case in particular of the spatial programming principle: physical space is treated as a tuple-space in which objects are automatically synchronized according to their spatial arrangement. We want to follow this approach by considering **richer and more expressive programming models.**

3.3. Acting on the environment

The conceptual tools we aim to study must be *frugal*: they use as less as possible resources, while having the possibility to use much more when it is required. Data needed by an application are not made available for "free"; for example, it costs energy to measure a characteristic of the environment, or to transmit it. So this "design frugality" requires **a fine-grained control** on how data is actually collected from the environment. The third research axis aims at designing solutions that give this control to application developers by **acting on the environment**.

Acting on the data collection. We want to be able to identify which information are really needed during the perception elaboration process. If a piece of data is missing to build a given information with the appropriate quality level, the data collection mechanism should find relevant information in the environment or modify the way it aggregates it. These could lead to a modification of the behavior of the network layer and the path the piece of data uses in the aggregation process.

Acting on object interactions. Objects in the environment could adapt their behavior in a way that strongly depends on the object itself and that is difficult to generalize. Beyond the specific behaviors of actuators triggered through specialized or standard interfaces, the production of information required by an application could necessitate an adaptation at the object level (eg. calibration, sampling). The environment should then be able to initiate such adaption transparently to the application, which may not know all objects it passes by.

Adapting object behaviors. The radio communication layers become more flexible and able to adapt the way they use energy to what is really required for a given transmission. We already study how beamforming technics could be used to adapt multicast strategy for video services. We want to show how playing with these new parameters of transmissions (eg. beamforming, power, ...) allows to control spatial relationships objects could have. There is a tradeoff to find between the capacity of the medium, the electromagnetic pollution and the reactivity of the environment. We plan to extend our previous work on interface selection and more generally on what we call **opportunistic networking**.

EVA Project-Team

3. Research Program

3.1. Pitch

Designing Tomorrow's Internet of (Important) Things

Inria-EVA is a leading research team in low-power wireless communications. The team pushes the limits of low-power wireless mesh networking by applying them to critical applications such as industrial control loops, with harsh reliability, scalability, security and energy constraints. Grounded in real-world use cases and experimentation, EVA co-chairs the IETF 6TiSCH and LAKE standardization working groups, co-leads Berkeley's OpenWSN project and works extensively with Analog Devices' SmartMesh IP networks. Inria-EVA is the birthplace of the Wattson Elements startup and the Falco solution. The team is associated with Prof. Glaser's (UC Berkeley) and Prof. Kerkez (U. Michigan) through the REALMS associate research team, and with OpenMote through a long-standing Memorandum of Understanding.

3.2. Physical Layer

We study how advanced physical layers can be used in low-power wireless networks. For instance, collaborative techniques such as multiple antennas (e.g. Massive MIMO technology) can improve communication efficiency. The core idea is to use massive network densification by drastically increasing the number of sensors in a given area in a Time Division Duplex (TDD) mode with time reversal. The first period allows the sensors to estimate the channel state and, after time reversal, the second period is to transmit the data sensed. Other techniques, such as interference cancellation, are also possible.

3.3. Wireless Access

Medium sharing in wireless systems has received substantial attention throughout the last decade. HiPERCOM2 has provided models to compare TDMA and CSMA. HiPERCOM2 has also studied how network nodes must be positioned to optimize the global throughput.

EVA pursues modeling tasks to compare access protocols, including multi-carrier access, adaptive CSMA (particularly in VANETs), as well as directional and multiple antennas. There is a strong need for determinism in industrial networks. The EVA team focuses particularly on scheduled medium access in the context of deterministic industrial networks; this involves optimizing the joint time slot and channel assignment. Distributed approaches are considered, and the EVA team determines their limits in terms of reliability, latency and throughput. Furthermore, adaptivity to application or environment changes are taken into account.

3.4. Coexistence of Wireless Technologies

Wireless technologies such as cellular, low-power mesh networks, (Low-Power) WiFi, and Bluetooth (low-energy) can reasonably claim to fit the requirements of the IoT. Each, however, uses different trade-offs between reliability, energy consumption and throughput. The EVA team studies the limits of each technology, and will develop clear criteria to evaluate which technology is best suited to a particular set of constraints.

Coexistence between these different technologies (or different deployments of the same technology in a common radio space) is a valid point of concern.

The EVA team aims at studying such coexistence, and, where necessary, propose techniques to improve it. Where applicable, the techniques will be put forward for standardization. Multiple technologies can also function in a symbiotic way.

For example, to improve the quality of experience provided to end users, a wireless mesh network can transport sensor and actuator data in place of a cellular network, when and where cellular connectivity is poor.

The EVA team studies how and when different technologies can complement one another. A specific example of a collaborative approach is Cognitive Radio Sensor Networks (CRSN).

3.5. Energy-Efficiency and Determinism

Reducing the energy consumption of low-power wireless devices remains a challenging task. The overall energy budget of a system can be reduced by using less power-hungry chips, and significant research is being done in that direction. That being said, power consumption is mostly influenced by the algorithms and protocols used in low-power wireless devices, since they influence the duty-cycle of the radio.

EVA will search for energy-efficient mechanisms in low-power wireless networks. One new requirement concerns the ability to predict energy consumption with a high degree of accuracy. Scheduled communication, such as the one used in the IEEE 802.15.4 TSCH (Time Slotted CHannel Hopping) standard, and by IETF 6TiSCH, allows for a very accurate prediction of the energy consumption of a chip. Power conservation will be a key issue in EVA.

To tackle this issue and match link-layer resources to application needs, EVA's 5-year research program dealing with Energy-Efficiency and Determinism centers around 3 studies:

- Performance Bounds of a TSCH network. We propose to study a low-power wireless TSCH network as a Networked Control System (NCS), and use results from the NCS literature. A large number of publications on NCS, although dealing with wireless systems, consider wireless links to have perfect reliability, and do not consider packet loss. Results from these papers can not therefore be applied directly to TSCH networks. Instead of following a purely mathematical approach to model the network, we propose to use a non-conventional approach and build an empirical model of a TSCH network.
- Distributed Scheduling in TSCH networks. Distributed scheduling is attractive due to its scalability and reactivity, but might result in a sub-optimal schedule. We continue this research by designing a distributed solution based on control theory, and verify how this solution can satisfy service level agreements in a dynamic environment.

3.6. Network Deployment

Since sensor networks are very often built to monitor geographical areas, sensor deployment is a key issue. The deployment of the network must ensure full/partial, permanent/intermittent coverage and connectivity. This technical issue leads to geometrical problems which are unusual in the networking domain.

We can identify two scenarios. In the first one, sensors are deployed over a given area to guarantee full coverage and connectivity, while minimizing the number of sensor nodes. In the second one, a network is re-deployed to improve its performance, possibly by increasing the number of points of interest covered, and by ensuring connectivity. EVA will investigate these two scenarios, as well as centralized and distributed approaches. The work starts with simple 2D models and will be enriched to take into account more realistic environment: obstacles, walls, 3D, fading.

3.7. Data Gathering and Dissemination

A large number of WSN applications mostly do data gathering (a.k.a "convergecast"). These applications usually require small delays for the data to reach the gateway node, requiring time consistency across gathered data. This time consistency is usually achieved by a short gathering period.

In many real WSN deployments, the channel used by the WSN usually encounters perturbations such as jamming, external interferences or noise caused by external sources (e.g. a polluting source such as a radar) or other coexisting wireless networks (e.g. WiFi, Bluetooth). Commercial sensor nodes can communicate on multiple frequencies as specified in the IEEE 802.15.4 standard. This reality has given birth to the multichannel communication paradigm in WSNs.

Multichannel WSNs significantly expand the capability of single-channel WSNs by allowing parallel transmissions, and avoiding congestion on channels or performance degradation caused by interfering devices.

In EVA, we will focus on raw data convergecast in multichannel low-power wireless networks. In this context, we are interested in centralized/distributed algorithms that jointly optimize the channel and time slot assignment used in a data gathering frame. The limits in terms of reliability, latency and bandwidth will be evaluated. Adaptivity to additional traffic demands will be improved.

3.8. Self-Learning Networks

To adapt to varying conditions in the environment and application requirements, the EVA team investigate self-learning networks. Machine learning approaches, based on experts and forecasters, are investigated to predict the quality of the wireless links in a WSN. This allows the routing protocol to avoid using links exhibiting poor quality and to change the route before a link failure. Additional applications include where to place the aggregation function in data gathering. In a content delivery network (CDN), it is very useful to predict popularity, expressed by the number of requests per day, for a multimedia content. The most popular contents are cached near the end-users to maximize the hit ratio of end-users' requests. Thus the satisfaction degree of end-users is maximized and the network overhead is minimized.

3.9. Internet of Things Security

Existing Internet threats might steal our digital information. Tomorrow's threats could disrupt power plants, home security systems, hospitals. The Internet of Things is bridging our digital security with personal safety. Popular magazines are full of stories of hacked devices (e.g. drone attack on Philips Hue), IoT botnets (e.g. Mirai), and inherent insecurity.

Why has the IoT industry failed to adopt the available computer security techniques and best practices? Our experience from research, industry collaborations, and the standards bodies has shown that the main challenges are:

1. The circumvention of the available technical solutions due to their inefficiency.
2. The lack of a user interface for configuring the product in the field resulting in default parameters being (re)used.
3. Poorly tested software, often lacking secure software upgrade mechanisms.

Our research goal is to contribute to a more secure IoT, by proposing technical solutions to these challenges for low-end IoT devices with immediate industrial applicability and transfer potential. We complement the existing techniques with the missing pieces to move towards truly usable and secure IoT systems.

FOCUS Project-Team

3. Research Program

3.1. Foundations 1: Models

The objective of Focus is to develop concepts, techniques, and possibly also tools, that may contribute to the analysis and synthesis of CBUS. Fundamental to these activities is *modeling*. Therefore designing, developing and studying computational models appropriate for CBUS is a central activity of the project. The models are used to formalise and verify important computational properties of the systems, as well as to propose new linguistic constructs.

The models we study are in the process calculi (e.g., the π-calculus) and λ-calculus tradition. Such models, with their emphasis on algebra, well address compositionality—a central property in our approach to problems. Accordingly, the techniques we employ are mainly operational techniques based on notions of behavioural equivalence, and techniques based on algebra, mathematical logics, and type theory.

3.2. Foundations 2: Foundational calculi and interaction

Modern distributed systems have witnessed a clear shift towards interaction and conversations as basic building blocks for software architects and programmers. The systems are made by components, that are supposed to interact and carry out dialogues in order to achieve some predefined goal; Web services are a good example of this. Process calculi are models that have been designed precisely with the goal of understanding interaction and composition. The theory and tools that have been developed on top of process calculi can set a basis with which CBUS challenges can be tackled. Indeed industrial proposals of languages for Web services such as BPEL are strongly inspired by process calculi, notably the π-calculus.

3.3. Foundations 3: Type systems and logics

Type systems and logics for reasoning on computations are among the most successful outcomes in the history of the research in λ-calculus and (more recently) in process calculi. Type systems can also represent a powerful means of specifying dialogues among components of CBUS. For instance—again referring to Web services—current languages for specifying interactions only express basic connectivity, ignoring causality and timing aspects (e.g., an intended order on the messages), and the alternative is to use Turing Complete languages that are however undecidable. Types can come at hand here: they can express causality and order information on messages [53], [49], [54], while remaining decidable systems.

3.4. Foundations 4: Implicit computational complexity

A number of elegant and powerful results have been recently obtained in implicit computational complexity in the λ-calculus in which ideas from Linear Logics enable a fine-grained control over computations. This experience can be profitable when tackling issues of CBUS related to resource consumption, such as resources allocation, access to resources, certification of bounds on resource consumption (e.g., ensuring that a service will answer to a request in time polynomial with respect to the size of the input data).

FUN Project-Team

3. Research Program

3.1. Introduction

We will focus on wireless ubiquitous networks that rely on constrained devices, i.e. with limited resources in terms of storage and computing capacities. They can be sensors, small robots, RFID readers or tags. A wireless sensor retrieves a physical measure such as light. A wireless robot is a wireless sensor that in addition has the ability to move by itself in a controlled way. A drone is a robot with the ability to manoeuvre in 3D (in the air or in the water). RFID tags are passive items that embed a unique identifier for a place or an object allowing accurate traceability. They can communicate only in the vicinity of an RFID reader. An RFID reader can be seen as a special kind of sensor in the network which data is the one read on tags. These devices may run on batteries that are not envisaged to be changed or recharged. These networks may be composed of ten to thousands of such heterogeneous devices for which energy is a key issue.

Today, most of these networks are homogeneous, i.e. composed of only one kind of devices. They have mainly been studied in application and technology silos. Because of this, they are approaching fundamental limitations especially in terms of topology deployment, management and communications, while exploiting the complementarity of heterogeneous devices and communication technologies would enlarge their capacities and the set of applications. Finally, these networks must work efficiently even in dynamic and realistic situations, i.e. they must consider by design the different dynamic parameters and automatically self-adapt to their variations.

Our overall goal is represented by Figure 1 . We will investigate wireless ubiquitous IoT services for constrained devices by smartly combining **different frequency bands** and **different medium access and routing techniques** over **heterogeneous devices** in a **distributed** and **opportunistic** fashion. Our approach will always deal with **hardware constraints** and take care of **security** and **energy** issues to provide protocols that ride on **synergy** and **self-organization** between devices.

The goal of the FUN project team is to provide these next generation networks with a set of innovative and distributed self-organizing cooperative protocols to raise them to a new level of scalability, autonomy, adaptability, manageability and performance. We aim to break these silos to exploit the full synergy between devices, making them cooperate in a single holistic network. We will consider them as networks of heterogeneous devices rather than a collection of heterogeneous networks.

To realize the full potential of these ubiquitous networks, there is a need to provide them with a set of tools that allow them to *(i)* (self-)deploy, *(ii)* self-organize, *(iii)* discover and locate each other, resources and services and *(iv)* communicate. These tools will be the basics for enabling cooperation, co-existence and witnessing a global efficient behavior. The deployment of these mechanisms is challenging since it should be achieved in spite of several limitations. The main difficulties are to provide such protocols in a **secured** and **energy-efficient** fashion in spite of:

- dynamic topology changes due to various factors such as the unreliability of the wireless medium, the wireless interferences between devices, node mobility and energy saving mechanisms;
- hardware constraints in terms of CPU and memory capacities that limit the operations and data each node can perform/collect;
- lacks of interoperability between applicative, hardware and technological silos that may prevent from data exchange between different devices.

3.1.1. Objectives and methodology

To reach our overall goal, we will pursue the two following objectives. These two objectives are orthogonal and can be carried on jointly:

1. Providing realistic complete self-organizing tools *e.g. vertical perspective*.
2. Going to heterogeneous energy-efficient performing wireless networks *e.g. horizontal perspective*.

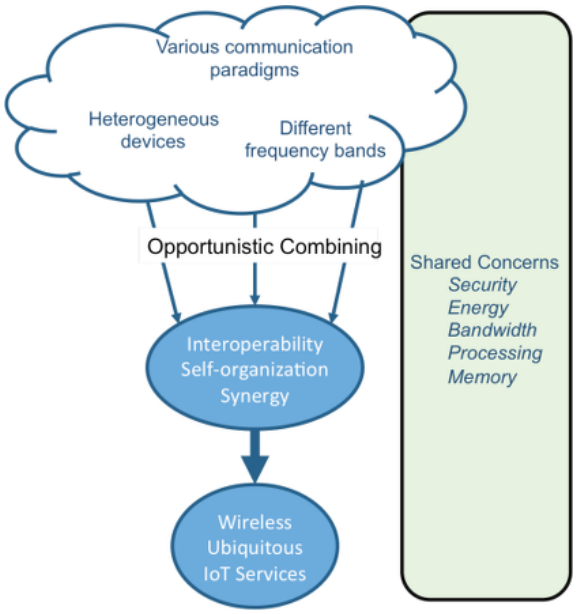


Figure 1. FUN's overal goal.

We give more details on these two objectives below. To achieve our main objectives, we will mainly apply the methodology depicted in Figure 2 combining both theoretical analysis and experimental validation. Mathematical tools will allow us to properly dimension a problem, formally define its limitations and needs to provide suitable protocols in response. Then, they will allow us to qualify the outcome solutions before we validate and stress them in real scenarios with regards to applications requirements. For this, we will realize proofs-of-concept with real scenarios and real devices. Differences between results and expectations will be analyzed in return in order to well understand them and integrate them by design for a better protocol self-adaptation capability.

3.2. Vertical Perspective

As mentioned, future ubiquitous networks evolve in dynamic and unpredictable environments. Also, they can be used in a large scope of applications that have several expectations in terms of performance and different contextual limitations. In this heterogeneous context, IoT devices must support multiple applications and relay traffic with non-deterministic pattern.

To make our solutions practical and efficient in real conditions, we will adopt the dual approach both *top-down* and *bottom-up*. The *top-down* approach will ensure that we consider the application (such as throughput, delay, energy consumption, etc.) and environmental limitations (such as deployment constraints, etc.). The *bottom-up* approach will ensure that we take account of the physical and hardware characteristics such as memory, CPU, energy capacities but also physical interferences and obstacles. With this integrated perpective, we will be in capacity to design well adapted **cross-layer** integrated protocols [59]. We will design jointly routing and MAC layers by taking dynamics occurring at the physical layer into account with a constant concern for energy and security. We will investigate new adaptive frequency hopping techniques combined with routing protocols [59], [45].

This vision will also allow us to integrate external factors by design in our protocols, in an opportunistic way. Yet, we will leverage on the occurrence of any of these phenomena rather than perceiving them as obstacles

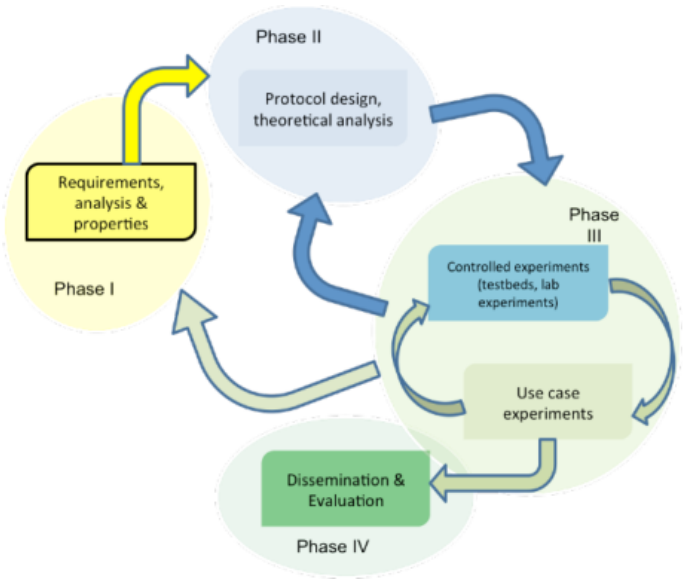


Figure 2. Methodology to be applied in FUN.

or limitations. As an example, we will rely on node undergone mobility to enhance routing performance as we have started to investigate in [55], [37]. On the same idea, when specific features are available like controlled mobility, we will exploit it to improve connectivity or coverage quality like in [49], [57], [31], [25].

3.3. Horizontal perspective

We aim at designing efficient tools for a plethora of wireless devices supporting highly heterogeneous technologies. We will thus investigate these networks from a horizontal perspective, e.g. by considering heterogeneity in low level communications layers.

Given the spectrum scarcity, they will probably need to coexist in the same frequency bands and sometimes for different purposes (RFID tag reading may use the same frequency bands as the wireless sensors). One important aspect to consider in this setting is how these different access technologies will interact with each other, and what are the mechanisms needed to be put in place to guarantee that all services obtain the required share of resources when needed. This problem appears in different application domains, ranging from traffic offloading to unlicensed bands by cellular networks and the need to coexist with WiFi and radars, from a scenario in which multiple-purpose IoT clouds coexist in a city [56]. We will thus explore the dynamics of these interactions and devise ways to ensure smooth coexistence while considering the heterogeneity of the devices involved, the access mechanisms used as well as the requirements of the services provided.

To face the spectrum scarcity, we will also investigate new alternative communication paradigms such as phonon-based or light-based communications as we have initiated in [42] and we will work on the coexistence of these technologies with traditional communication techniques, specifically by investigating efficient switching techniques from one communication technology to the other (they were most focused on the security aspects, to prevent jamming attacks). Resilience and reliability of the whole system will be the key factors to be taken into account [43], [39], [18].

As a more prospective activity, we consider exploring software and communication security for IoT. This is challenging given that existing solutions do not address systems that are both constrained and networked [44].

Finally, in order to contribute to a better interoperability between all these technologies, we will continue to contribute to standardization bodies such as IETF and EPC Global.

GANG Project-Team

3. Research Program

3.1. Graph and Combinatorial Algorithms

We focus on two approaches for designing algorithms for large graphs: decomposing the graph and relying on simple graph traversals.

3.1.1. Graph Search

We more deeply study multi-sweep graph searches. In this domain a graph search only yields a total ordering of the vertices which can be used by the subsequent graph searches. This technique can be used on huge graphs and does not need extra memory. We have already obtained preliminary results in this direction and many well-known graph algorithms can be put in this framework. The idea behind this approach is that each sweep discovers some structure of the graph. At the end of the process either we have found the underlying structure (for example an interval representation for an interval graph) or an approximation of it (for example in hard discrete optimization problems). We envision applications to exact computations of centers in huge graphs, to underlying combinatorial optimization problems, but also to networks arising in biology.

3.1.2. Graph Decomposition

In order to summarize a graph into a more compact and more human-readable form, we introduced the hub-laminar decomposition. It is suitable for graphs that are dominated by long isometric cycles or shortest paths, called laminar, which meet only at their extremities, called hubs. Computing this decomposition is NP-hard but a canonical approximation may be computed under some hypotheses on the distances between hubs. It provides a distance labelling for the decomposable graphs. We also investigated the case were the decomposition is reduced to a single cycle, yielding the problem of fining the longest isometric cycle, which is NP-complete and for which a first approximation algorithm was proposed in ENTCS.

3.1.3. Graph Exploration

In the course of graph exploration, a mobile agent is expected to regularly visit all the nodes of an unknown network, trying to discover all its nodes as quickly as possible. Our research focuses on the design and analysis of agent-based algorithms for exploration-type problems, which operate efficiently in a dynamic network environment, and satisfy imposed constraints on local computational resources, performance, and resilience. Our recent contributions in this area concern the design of fast deterministic algorithms for teams of agents operating in parallel in a graph, with limited or no persistent state information available at nodes. We plan further studies to better understand the impact of memory constraints and of the availability of true randomness on efficiency of the graph exploration process.

3.2. Distributed Computing

The distributed computing community can be viewed as a union of two sub-communities. This is also true in our team. Although they have interactions, they are disjoint enough not to leverage each other's results. At a high level, one is mostly interested in timing issues (clock drifts, link delays, crashes, etc.) while the other one is mostly interested in spatial issues (network structure, memory requirements, etc.). Indeed, one sub-community is mostly focusing on the combined impact of asynchronism and faults on distributed computation, while the other addresses the impact of network structural properties on distributed computation. Both communities address various forms of computational complexity, through the analysis of different concepts. This includes, e.g., failure detectors and wait-free hierarchy for the former community and compact labeling schemes, and computing with advice for the latter community. We have an ambitious project to achieve the reconciliation between the two communities by focusing on the same class of problems, the yes/no-problems, and establishing the scientific foundations for building up a consistent theory of computability and complexity

for distributed computing. The main question addressed is therefore: is the absence of globally coherent computational complexity theories covering more than fragments of distributed computing, inherent to the field? One issue is obviously the types of problems located at the core of distributed computing. Tasks like consensus, leader election, and broadcasting are of very different nature. They are not *yes-no* problems, neither are they minimization problems. Coloring and Minimal Spanning Tree are optimization problems but we are often more interested in constructing an optimal solution than in verifying the correctness of a given solution. Still, it makes full sense to analyze the *yes-no* problems corresponding to checking the validity of the output of tasks. Another issue is the power of individual computation. The FLP impossibility result as well as Linial's lower bound hold independently of the individual computational power of the involved computing entities. For instance, the individual power of solving NP-hard problems in constant time would not help overcoming these limits, which are inherent to the fact that computation is distributed. A third issue is the abundance of models for distributed computing frameworks, from shared memory to message passing, spanning all kinds of specific network structures (complete graphs, unit-disk graphs, etc.) and/or timing constraints (from complete synchronism to full asynchronism). There are however models, typically the wait-free model and the LOCAL model, which, though they do not claim to reflect accurately real distributed computing systems, enable focusing on some core issues. Our ongoing research program is to carry many important notions of Distributed Computing into a *standard* computational complexity.

3.3. Network Algorithms and Analysis

Based on our scientific expertise in both graph algorithms and distributed algorithms, we plan to analyze the behavior of various networks such as future Internet, social networks, overlay networks resulting from distributed applications or online social networks.

3.3.1. Information Dissemination

One of the key aspects of networks resides in the dissemination of information among the nodes. We aim at analyzing various procedures of information propagation from dedicated algorithms to simple distributed schemes such as flooding. We also consider various models, e.g. where noise can alter information as it propagates or where memory of nodes is limited.

3.3.2. Routing Paradigms

We try to explore new routing paradigms such as greedy routing in social networks for example. We are also interested in content centric networking where routing is based on content name rather than content address. One of our targets is multiple path routing: how to design forwarding tables providing multiple disjoint paths to the destination?

3.3.3. Beyond Peer-to-Peer

Based on our past experience of peer-to-peer application design, we would like to broaden the spectrum of distributed applications where new efficient algorithms can be designed and their analysis can be performed. We especially target online social networks as we see them as collaborative tools for exchanging information. A basic question resides in making the right connections for gathering filtered and accurate information with sufficient coverage.

3.3.4. SAT and Forwarding Information Verification

As forwarding tables of networks grow and are sometimes manually modified, the problem of verifying them becomes critical and has recently gained interest. Some problems that arise in network verification such as loop detection for example, may be naturally encoded as Boolean Satisfiability problems. Beside theoretical interest in complexity proofs, this encoding allows one to solve these problems by taking advantage of efficient Satisfiability testing solvers. Indeed, SAT solvers have proved to be very efficient in solving problems coming from various areas (Circuit Verification, Dependency and Conflicts in Software distributions...) and encoded in Conjunctive Normal Form. To test an approach using SAT solvers in network verification, one needs to collect data sets from a real network and to develop good models for generating realistic networks. The technique of encoding and the solvers themselves need to be adapted to this kind of problems. All this represents a rich experimental field of future research.

3.3.5. Network Analysis

Finally, we are interested in analyzing the structural properties of practical networks. This can include diameter computation or ranking of nodes. As we mostly consider large networks, we are often interested in efficient heuristics. Ideally, we target heuristics that give exact answers and are reasonably fast in practice although short computation time is not guaranteed for all networks. We have already designed such heuristics for diameter computation; understanding the structural properties that enable short computation time in practice is still an open question.

3.3.6. Network Parameter

Betweenness centrality is a graph parameter that has been successfully applied to network analysis. In the context of computer networks, it was considered for various objectives, ranging from routing to service placement. However, as observed by Maccari et al. [INFOCOM 2018], research on betweenness centrality for improving protocols was hampered by the lack of a usable, fully distributed algorithm for computing this parameter. In [21], we resolved this issue by designing an efficient algorithm for computing betweenness centrality, which can be implemented by minimal modifications to any distance-vector routing protocol based on Bellman-Ford. The convergence time of our implementation is shown to be proportional to the diameter of the network.

HIEPACS Project-Team

3. Research Program

3.1. Introduction

The methodological component of HIEPACS concerns the expertise for the design as well as the efficient and scalable implementation of highly parallel numerical algorithms to perform frontier simulations. In order to address these computational challenges a hierarchical organization of the research is considered. In this bottom-up approach, we first consider in Section 3.2 generic topics concerning high performance computational science. The activities described in this section are transversal to the overall project and their outcome will support all the other research activities at various levels in order to ensure the parallel scalability of the algorithms. The aim of this activity is not to study general purpose solution but rather to address these problems in close relation with specialists of the field in order to adapt and tune advanced approaches in our algorithmic designs. The next activity, described in Section 3.3 , is related to the study of parallel linear algebra techniques that currently appear as promising approaches to tackle huge problems on extreme scale platforms. We highlight the linear problems (linear systems or eigenproblems) because they are in many large scale applications the main computational intensive numerical kernels and often the main performance bottleneck. These parallel numerical techniques will be the basis of both academic and industrial collaborations, some are described in Section 4.1 , but will also be closely related to some functionalities developed in the parallel fast multipole activity described in Section 3.4 . Finally, as the accuracy of the physical models increases, there is a real need to go for parallel efficient algorithm implementation for multiphysics and multiscale modeling in particular in the context of code coupling. The challenges associated with this activity will be addressed in the framework of the activity described in Section 3.5 .

Currently, we have one major application (see Section 4.1) that is in material physics. We will collaborate to all steps of the design of the parallel simulation tool. More precisely, our applied mathematics skill will contribute to the modelling, our advanced numerical schemes will help in the design and efficient software implementation for very large parallel simulations. We also participate to a few co-design actions in close collaboration with some applicative groups. The objective of this activity is to instantiate our expertise in fields where they are critical for designing scalable simulation tools. We refer to Section 4.2 for a detailed description of these activities.

3.2. High-performance computing on next generation architectures

Participants: Emmanuel Agullo, Olivier Beaumont, Olivier Coulaud, Pierre Esterie, Lionel Eyraud-Dubois, Mathieu Faverge, Luc Giraud, Abdou Guermouche, Gilles Marait, Pierre Ramet, Jean Roman, Nick Schenkels, Alena Shilova, Mathieu Verite.

The research directions proposed in HIEPACS are strongly influenced by both the applications we are studying and the architectures that we target (i.e., massively parallel heterogeneous many-core architectures, ...). Our main goal is to study the methodology needed to efficiently exploit the new generation of high-performance computers with all the constraints that it induces. To achieve this high-performance with complex applications we have to study both algorithmic problems and the impact of the architectures on the algorithm design.

From the application point of view, the project will be interested in multiresolution, multiscale and hierarchical approaches which lead to multi-level parallelism schemes. This hierarchical parallelism approach is necessary to achieve good performance and high-scalability on modern massively parallel platforms. In this context, more specific algorithmic problems are very important to obtain high performance. Indeed, the kind of applications we are interested in are often based on data redistribution for example (e.g., code coupling applications). This well-known issue becomes very challenging with the increase of both the number of computational nodes and the amount of data. Thus, we have both to study new algorithms and to adapt the

existing ones. In addition, some issues like task scheduling have to be restudied in this new context. It is important to note that the work developed in this area will be applied for example in the context of code coupling (see Section 3.5).

Considering the complexity of modern architectures like massively parallel architectures or new generation heterogeneous multicore architectures, task scheduling becomes a challenging problem which is central to obtain a high efficiency. With the recent addition of colleagues from the scheduling community (O. Beaumont and L. Eyraud-Dubois), the team is better equipped than ever to design scheduling algorithms and models specifically tailored to our target problems. It is important to note that this topic is strongly linked to the underlying programming model. Indeed, considering multicore and heterogeneous architectures, it has appeared, in the last five years, that the best programming model is an approach mixing multi-threading within computational nodes and message passing between them. In the last five years, a lot of work has been developed in the high-performance computing community to understand what is critic to efficiently exploit massively multicore platforms that will appear in the near future. It appeared that the key for the performance is firstly the granularity of the computations. Indeed, in such platforms the granularity of the parallelism must be small so that we can feed all the computing units with a sufficient amount of work. It is thus very crucial for us to design new high performance tools for scientific computing in this new context. This will be developed in the context of our solvers, for example, to adapt to this new parallel scheme. Secondly, the larger the number of cores inside a node, the more complex the memory hierarchy. This remark impacts the behavior of the algorithms within the node. Indeed, on this kind of platforms, NUMA effects will be more and more problematic. Thus, it is very important to study and design data-aware algorithms which take into account the affinity between computational threads and the data they access. This is particularly important in the context of our high-performance tools. Note that this work has to be based on an intelligent cooperative underlying run-time (like the tools developed by the Inria STORM Project-Team) which allows a fine management of data distribution within a node.

Another very important issue concerns high-performance computing using “heterogeneous” resources within a computational node. Indeed, with the deployment of the GPU and the use of more specific co-processors, it is important for our algorithms to efficiently exploit these new type of architectures. To adapt our algorithms and tools to these accelerators, we need to identify what can be done on the GPU for example and what cannot. Note that recent results in the field have shown the interest of using both regular cores and GPU to perform computations. Note also that in opposition to the case of the parallelism granularity needed by regular multicore architectures, GPU requires coarser grain parallelism. Thus, making both GPU and regular cores work all together will lead to two types of tasks in terms of granularity. This represents a challenging problem especially in terms of scheduling. From this perspective, we investigate new approaches for composing parallel applications within a runtime system for heterogeneous platforms.

In the context of scaling up, and particularly in the context of minimizing energy consumption, it is generally acknowledged that the solution lies in the use of heterogeneous architectures, where each resource is particularly suited to specific types of tasks, and in a fine control at the algorithmic level of data movements and the trade-offs to be made between computation and communication. In this context, we are particularly interested in the optimization of the training phase of deep convolutional neural networks which consumes a lot of memory and for which it is possible to exchange computations for data movements and memory occupation. We are also interested in the complexity introduced by resource heterogeneity itself, both from a theoretical point of view on the complexity of scheduling problems and from a more practical point of view on the implementation of specific kernels in dense or sparse linear algebra.

In order to achieve an advanced knowledge concerning the design of efficient computational kernels to be used on our high performance algorithms and codes, we will develop research activities first on regular frameworks before extending them to more irregular and complex situations. In particular, we will work first on optimized dense linear algebra kernels and we will use them in our more complicated direct and hybrid solvers for sparse linear algebra and in our fast multipole algorithms for interaction computations. In this context, we will participate to the development of those kernels in collaboration with groups specialized in dense linear algebra. In particular, we intend develop a strong collaboration with the group of Jack Dongarra

at the University of Tennessee and collaborating research groups. The objectives will be to develop dense linear algebra algorithms and libraries for multicore architectures in the context the PLASMA project and for GPU and hybrid multicore/GPU architectures in the context of the MAGMA project. A new solver has emerged from the associate team, Chameleon. While PLASMA and MAGMA focus on multicore and GPU architectures, respectively, Chameleon makes the most out of heterogeneous architectures thanks to task-based dynamic runtime systems.

A more prospective objective is to study the resiliency in the context of large-scale scientific applications for massively parallel architectures. Indeed, with the increase of the number of computational cores per node, the probability of a hardware crash on a core or of a memory corruption is dramatically increased. This represents a crucial problem that needs to be addressed. However, we will only study it at the algorithmic/application level even if it needed lower-level mechanisms (at OS level or even hardware level). Of course, this work can be performed at lower levels (at operating system) level for example but we do believe that handling faults at the application level provides more knowledge about what has to be done (at application level we know what is critical and what is not). The approach that we will follow will be based on the use of a combination of fault-tolerant implementations of the run-time environments we use (like for example ULFM) and an adaptation of our algorithms to try to manage this kind of faults. This topic represents a very long range objective which needs to be addressed to guaranty the robustness of our solvers and applications.

Finally, it is important to note that the main goal of HIEPACS is to design tools and algorithms that will be used within complex simulation frameworks on next-generation parallel machines. Thus, we intend with our partners to use the proposed approach in complex scientific codes and to validate them within very large scale simulations as well as designing parallel solution in co-design collaborations.

3.3. High performance solvers for large linear algebra problems

Participants: Emmanuel Agullo, Olivier Coulaud, Tony Delarue, Mathieu Faverge, Aurélien Falco, Marek Felsoci, Luc Giraud, Abdou Guermouche, Esragul Korkmaz, Gilles Marait, Van Gia Thinh Nguyen, Jean Rene Poirier, Pierre Ramet, Jean Roman, Cristobal Samaniego Alvarado, Guillaume Sylvand, Nicolas Venkovic, Yanfei Xiang.

Starting with the developments of basic linear algebra kernels tuned for various classes of computers, a significant knowledge on the basic concepts for implementations on high-performance scientific computers has been accumulated. Further knowledge has been acquired through the design of more sophisticated linear algebra algorithms fully exploiting those basic intensive computational kernels. In that context, we still look at the development of new computing platforms and their associated programming tools. This enables us to identify the possible bottlenecks of new computer architectures (memory path, various level of caches, inter processor or node network) and to propose ways to overcome them in algorithmic design. With the goal of designing efficient scalable linear algebra solvers for large scale applications, various tracks will be followed in order to investigate different complementary approaches. Sparse direct solvers have been for years the methods of choice for solving linear systems of equations, it is nowadays admitted that classical approaches are not scalable neither from a computational complexity nor from a memory view point for large problems such as those arising from the discretization of large 3D PDE problems. We will continue to work on sparse direct solvers on the one hand to make sure they fully benefit from most advanced computing platforms and on the other hand to attempt to reduce their memory and computational costs for some classes of problems where data sparse ideas can be considered. Furthermore, sparse direct solvers are a key building boxes for the design of some of our parallel algorithms such as the hybrid solvers described in the sequel of this section. Our activities in that context will mainly address preconditioned Krylov subspace methods; both components, preconditioner and Krylov solvers, will be investigated. In this framework, and possibly in relation with the research activity on fast multipole, we intend to study how emerging $\mathcal{H}$-matrix arithmetic can benefit to our solver research efforts.

3.3.1. Parallel sparse direct solvers

For the solution of large sparse linear systems, we design numerical schemes and software packages for direct and hybrid parallel solvers. Sparse direct solvers are mandatory when the linear system is very ill-conditioned; such a situation is often encountered in structural mechanics codes, for example. Therefore, to obtain an industrial software tool that must be robust and versatile, high-performance sparse direct solvers are mandatory, and parallelism is then necessary for reasons of memory capability and acceptable solution time. Moreover, in order to solve efficiently 3D problems with more than 50 million unknowns, which is now a reachable challenge with new multicore supercomputers, we must achieve good scalability in time and control memory overhead. Solving a sparse linear system by a direct method is generally a highly irregular problem that induces some challenging algorithmic problems and requires a sophisticated implementation scheme in order to fully exploit the capabilities of modern supercomputers.

New supercomputers incorporate many microprocessors which are composed of one or many computational cores. These new architectures induce strongly hierarchical topologies. These are called NUMA architectures. In the context of distributed NUMA architectures, in collaboration with the Inria STORM team, we study optimization strategies to improve the scheduling of communications, threads and I/O. We have developed dynamic scheduling designed for NUMA architectures in the PaStiX solver. The data structures of the solver, as well as the patterns of communication have been modified to meet the needs of these architectures and dynamic scheduling. We are also interested in the dynamic adaptation of the computation grain to use efficiently multi-core architectures and shared memory. Experiments on several numerical test cases have been performed to prove the efficiency of the approach on different architectures. Sparse direct solvers such as PaStiX are currently limited by their memory requirements and computational cost. They are competitive for small matrices but are often less efficient than iterative methods for large matrices in terms of memory. We are currently accelerating the dense algebra components of direct solvers using block low-rank compression techniques.

In collaboration with the ICL team from the University of Tennessee, and the STORM team from Inria, we are evaluating the way to replace the embedded scheduling driver of the PaStiX solver by one of the generic frameworks, PaRSEC or StarPU, to execute the task graph corresponding to a sparse factorization. The aim is to design algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computer equipped with GPU accelerators. More generally, this work will be performed in the context of the ANR SOLHARIS project which aims at designing high performance sparse direct solvers for modern heterogeneous systems. This ANR project involves several groups working either on the sparse linear solver aspects (HIEPACS and ROMA from Inria and APO from IRIT), on runtime systems (STORM from Inria) or scheduling algorithms (HIEPACS and ROMA from Inria). The results of these efforts will be validated in the applications provided by the industrial project members, namely CEA-CESTA and Airbus Central R & T.

3.3.2. Hybrid direct/iterative solvers based on algebraic domain decomposition techniques

One route to the parallel scalable solution of large sparse linear systems in parallel scientific computing is the use of hybrid methods that hierarchically combine direct and iterative methods. These techniques inherit the advantages of each approach, namely the limited amount of memory and natural parallelization for the iterative component and the numerical robustness of the direct part. The general underlying ideas are not new since they have been intensively used to design domain decomposition techniques; those approaches cover a fairly large range of computing techniques for the numerical solution of partial differential equations (PDEs) in time and space. Generally speaking, it refers to the splitting of the computational domain into sub-domains with or without overlap. The splitting strategy is generally governed by various constraints/objectives but the main one is to express parallelism. The numerical properties of the PDEs to be solved are usually intensively exploited at the continuous or discrete levels to design the numerical algorithms so that the resulting specialized technique will only work for the class of linear systems associated with the targeted PDE.

In that context, we continue our effort on the design of algebraic non-overlapping domain decomposition techniques that rely on the solution of a Schur complement system defined on the interface introduced by the partitioning of the adjacency graph of the sparse matrix associated with the linear system. Although it is better conditioned than the original system the Schur complement needs to be precondition to be amenable to

a solution using a Krylov subspace method. Different hierarchical preconditioners will be considered, possibly multilevel, to improve the numerical behaviour of the current approaches implemented in our software library MaPHyS. This activity will be developed further developped in the H2020 EoCoE2 project. In addition to this numerical studies, advanced parallel implementation will be developed that will involve close collaborations between the hybrid and sparse direct activities.

3.3.3. Linear Krylov solvers

Preconditioning is the main focus of the two activities described above. They aim at speeding up the convergence of a Krylov subspace method that is the complementary component involved in the solvers of interest for us. In that framework, we believe that various aspects deserve to be investigated; we will consider the following ones:

- preconditioned block Krylov solvers for multiple right-hand sides. In many large scientific and industrial applications, one has to solve a sequence of linear systems with several right-hand sides given simultaneously or in sequence (radar cross section calculation in electromagnetism, various source locations in seismic, parametric studies in general, ...). For “simultaneous" right-hand sides, the solvers of choice have been for years based on matrix factorizations as the factorization is performed once and simple and cheap block forward/backward substitutions are then performed. In order to effectively propose alternative to such solvers, we need to have efficient preconditioned Krylov subspace solvers. In that framework, block Krylov approaches, where the Krylov spaces associated with each right-hand side are shared to enlarge the search space will be considered. They are not only attractive because of this numerical feature (larger search space), but also from an implementation point of view. Their block-structures exhibit nice features with respect to data locality and re-usability that comply with the memory constraint of multicore architectures. We will continue the numerical study and design of the block GMRES variant that combines inexact breakdown detection, deflation at restart and subspace recycling. Beyond new numerical investigations, a software implementation to be included in our linear solver libray Fabulous originately developed in the context of the DGA HiBox project and further developped in the LynCs (Linear Algebra, Krylov-subspace methods, and multi-grid solvers for the discovery of New Physics) sub-project of Prace-6IP.
- Extension or modification of Krylov subspace algorithms for multicore architectures: finally to match as much as possible to the computer architecture evolution and get as much as possible performance out of the computer, a particular attention will be paid to adapt, extend or develop numerical schemes that comply with the efficiency constraints associated with the available computers. Nowadays, multicore architectures seem to become widely used, where memory latency and bandwidth are the main bottlenecks; investigations on communication avoiding techniques will be undertaken in the framework of preconditioned Krylov subspace solvers as a general guideline for all the items mentioned above.

3.3.4. Eigensolvers

Many eigensolvers also rely on Krylov subspace techniques. Naturally some links exist between the Krylov subspace linear solvers and the Krylov subspace eigensolvers. We plan to study the computation of eigenvalue problems with respect to the following two different axes:

- Exploiting the link between Krylov subspace methods for linear system solution and eigensolvers, we intend to develop advanced iterative linear methods based on Krylov subspace methods that use some spectral information to build part of a subspace to be recycled, either though space augmentation or through preconditioner update. This spectral information may correspond to a certain part of the spectrum of the original large matrix or to some approximations of the eigenvalues obtained by solving a reduced eigenproblem. This technique will also be investigated in the framework of block Krylov subspace methods.
- In the context of the calculation of the ground state of an atomistic system, eigenvalue computation is a critical step; more accurate and more efficient parallel and scalable eigensolvers are required.

3.3.5. Fast Solvers for FEM/BEM Coupling

In this research project, we are interested in the design of new advanced techniques to solve large mixed dense/sparse linear systems, the extensive comparison of these new approaches to the existing ones, and the application of these innovative ideas on realistic industrial test cases in the domain of aeroacoustics (in collaboration with Airbus Central R & T).

- The use of $\mathcal{H}$-matrix solvers on these problems has been investigated in the context of the PhD of A. Falco. Airbus CR&T, in collaboration with Inria Bordeaux Sud-Ouest, has developed a task-based $\mathcal{H}$-matrix solver on top of the runtime engine StarPU. Ideas coming from the field of sparse direct solvers (such as nested dissection or symbolic factorization) have been tested within $\mathcal{H}$-matrices.
- The question of parallel scalability of task-based tools is an active subject of research, using new communication engine such as NewMadeleine, that will be investigated during this project, in conjunction with new algorithmic ideas on the task-based writing of $\mathcal{H}$-matrix algorithms.
- Naturally, comparison with existing tools will be performed on large realistic test cases. Coupling schemes between these tools and the hierarchical methods used in $\mathcal{H}$-matrix will be developed and benched as well.

3.4. High performance Fast Multipole Method for N-body problems

Participants: Emmanuel Agullo, Olivier Coulaud, Pierre Esterie, Guillaume Sylvand.

In most scientific computing applications considered nowadays as computational challenges (like biological and material systems, astrophysics or electromagnetism), the introduction of hierarchical methods based on an octree structure has dramatically reduced the amount of computation needed to simulate those systems for a given accuracy. For instance, in the N-body problem arising from these application fields, we must compute all pairwise interactions among N objects (particles, lines, ...) at every timestep. Among these methods, the Fast Multipole Method (FMM) developed for gravitational potentials in astrophysics and for electrostatic (coulombic) potentials in molecular simulations solves this N-body problem for any given precision with $O(N)$ runtime complexity against $O(N^2)$ for the direct computation.

The potential field is decomposed in a near field part, directly computed, and a far field part approximated thanks to multipole and local expansions. We introduced a matrix formulation of the FMM that exploits the cache hierarchy on a processor through the Basic Linear Algebra Subprograms (BLAS). Moreover, we developed a parallel adaptive version of the FMM algorithm for heterogeneous particle distributions, which is very efficient on parallel clusters of SMP nodes. Finally on such computers, we developed the first hybrid MPI-thread algorithm, which enables to reach better parallel efficiency and better memory scalability. We plan to work on the following points in HIEPACS .

3.4.1. Improvement of calculation efficiency

Nowadays, the high performance computing community is examining alternative architectures that address the limitations of modern cache-based designs. GPU (Graphics Processing Units) and the Cell processor have thus already been used in astrophysics and in molecular dynamics. The Fast Mutipole Method has also been implemented on GPU. We intend to examine the potential of using these forthcoming processors as a building block for high-end parallel computing in N-body calculations. More precisely, we want to take advantage of our specific underlying BLAS routines to obtain an efficient and easily portable FMM for these new architectures. Algorithmic issues such as dynamic load balancing among heterogeneous cores will also have to be solved in order to gather all the available computation power. This research action will be conduced on close connection with the activity described in Section 3.2 .

3.4.2. Non uniform distributions

In many applications arising from material physics or astrophysics, the distribution of the data is highly non uniform and the data can grow between two time steps. As mentioned previously, we have proposed a hybrid MPI-thread algorithm to exploit the data locality within each node. We plan to further improve the load

balancing for highly non uniform particle distributions with small computation grain thanks to dynamic load balancing at the thread level and thanks to a load balancing correction over several simulation time steps at the process level.

3.4.3. Fast multipole method for dislocation operators

The engine that we develop will be extended to new potentials arising from material physics such as those used in dislocation simulations. The interaction between dislocations is long ranged ($O(1/r)$) and anisotropic, leading to severe computational challenges for large-scale simulations. Several approaches based on the FMM or based on spatial decomposition in boxes are proposed to speed-up the computation. In dislocation codes, the calculation of the interaction forces between dislocations is still the most CPU time consuming. This computation has to be improved to obtain faster and more accurate simulations. Moreover, in such simulations, the number of dislocations grows while the phenomenon occurs and these dislocations are not uniformly distributed in the domain. This means that strategies to dynamically balance the computational load are crucial to achieve high performance.

3.4.4. Fast multipole method for boundary element methods

The boundary element method (BEM) is a well known solution of boundary value problems appearing in various fields of physics. With this approach, we only have to solve an integral equation on the boundary. This implies an interaction that decreases in space, but results in the solution of a dense linear system with $O(N^3)$ complexity. The FMM calculation that performs the matrix-vector product enables the use of Krylov subspace methods. Based on the parallel data distribution of the underlying octree implemented to perform the FMM, parallel preconditioners can be designed that exploit the local interaction matrices computed at the finest level of the octree. This research action will be conduced on close connection with the activity described in Section 3.3 . Following our earlier experience, we plan to first consider approximate inverse preconditionners that can efficiently exploit these data structures.

3.5. Load balancing algorithms for complex simulations

Participants: Cyril Bordage, Aurélien Esnard, Pierre Ramet.

Many important physical phenomena in material physics and climatology are inherently complex applications. They often use multi-physics or multi-scale approaches, which couple different models and codes. The key idea is to reuse available legacy codes through a coupling framework instead of merging them into a stand-alone application. There is typically one model per different scale or physics and each model is implemented by a parallel code.

For instance, to model a crack propagation, one uses a molecular dynamic code to represent the atomistic scale and an elasticity code using a finite element method to represent the continuum scale. Indeed, fully microscopic simulations of most domains of interest are not computationally feasible. Combining such different scales or physics is still a challenge to reach high performance and scalability.

Another prominent example is found in the field of aeronautic propulsion: the conjugate heat transfer simulation in complex geometries (as developed by the CFD team of CERFACS) requires to couple a fluid/convection solver (AVBP) with a solid/conduction solver (AVTP). As the AVBP code is much more CPU consuming than the AVTP code, there is an important computational imbalance between the two solvers.

In this context, one crucial issue is undoubtedly the load balancing of the whole coupled simulation that remains an open question. The goal here is to find the best data distribution for the whole coupled simulation and not only for each stand-alone code, as it is most usually done. Indeed, the naive balancing of each code on its own can lead to an important imbalance and to a communication bottleneck during the coupling phase, which can drastically decrease the overall performance. Therefore, we argue that it is required to model the coupling itself in order to ensure a good scalability, especially when running on massively parallel architectures (tens of thousands of processors/cores). In other words, one must develop new algorithms and software implementation to perform a *coupling-aware* partitioning of the whole application. Another related problem is the problem of resource allocation. This is particularly important for the global coupling efficiency and

scalability, because each code involved in the coupling can be more or less computationally intensive, and there is a good trade-off to find between resources assigned to each code to avoid that one of them waits for the other(s). What does furthermore happen if the load of one code dynamically changes relatively to the other one? In such a case, it could be convenient to dynamically adapt the number of resources used during the execution.

There are several open algorithmic problems that we investigate in the HIEPACS project-team. All these problems uses a similar methodology based upon the graph model and are expressed as variants of the classic graph partitioning problem, using additional constraints or different objectives.

3.5.1. Dynamic load-balancing with variable number of processors

As a preliminary step related to the dynamic load balancing of coupled codes, we focus on the problem of dynamic load balancing of a single parallel code, with variable number of processors. Indeed, if the workload varies drastically during the simulation, the load must be redistributed regularly among the processors. Dynamic load balancing is a well studied subject but most studies are limited to an initially fixed number of processors. Adjusting the number of processors at runtime allows one to preserve the parallel code efficiency or keep running the simulation when the current memory resources are exceeded. We call this problem, *MxN graph repartitioning*.

We propose some methods based on graph repartitioning in order to re-balance the load while changing the number of processors. These methods are split in two main steps. Firstly, we study the migration phase and we build a “good” migration matrix minimizing several metrics like the migration volume or the number of exchanged messages. Secondly, we use graph partitioning heuristics to compute a new distribution optimizing the migration according to the previous step results.

3.5.2. Load balancing of coupled codes

As stated above, the load balancing of coupled code is a major issue, that determines the performance of the complex simulation, and reaching high performance can be a great challenge. In this context, we develop new graph partitioning techniques, called *co-partitioning*. They address the problem of load balancing for two coupled codes: the key idea is to perform a "coupling-aware" partitioning, instead of partitioning these codes independently, as it is classically done. More precisely, we propose to enrich the classic graph model with *inter-edges*, which represent the coupled code interactions. We describe two new algorithms, and compare them to the naive approach. In the preliminary experiments we perform on synthetically-generated graphs, we notice that our algorithms succeed to balance the computational load in the coupling phase and in some cases they succeed to reduce the coupling communications costs. Surprisingly, we notice that our algorithms do not degrade significantly the global graph edge-cut, despite the additional constraints that they impose.

Besides this, our co-partitioning technique requires to use graph partitioning with *fixed vertices*, that raises serious issues with state-of-the-art software, that are classically based on the well-known recursive bisection paradigm (RB). Indeed, the RB method often fails to produce partitions of good quality. To overcome this issue, we propose a *new* direct k-way greedy graph growing algorithm, called KGGGP, that overcomes this issue and succeeds to produce partition with better quality than RB while respecting the constraint of fixed vertices. Experimental results compare KGGGP against state-of-the-art methods, such as Scotch, for real-life graphs available from the popular *DIMACS'10* collection.

3.5.3. Load balancing strategies for hybrid sparse linear solvers

Graph handling and partitioning play a central role in the activity described here but also in other numerical techniques detailed in sparse linear algebra Section. The Nested Dissection is now a well-known heuristic for sparse matrix ordering to both reduce the fill-in during numerical factorization and to maximize the number of independent computation tasks. By using the block data structure induced by the partition of separators of the original graph, very efficient parallel block solvers have been designed and implemented according to super-nodal or multi-frontal approaches. Considering hybrid methods mixing both direct and iterative solvers such as MaPHyS, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the size of interfaces is a key point for load balancing and efficiency in a parallel context.

We intend to revisit some well-known graph partitioning techniques in the light of the hybrid solvers and design new algorithms to be tested in the Scotch package.

INDES Project-Team

3. Research Program

3.1. Parallelism, concurrency, and distribution

Concurrency management is at the heart of diffuse programming. Since the execution platforms are highly heterogeneous, many different concurrency principles and models may be involved. Asynchronous concurrency is the basis of shared-memory process handling within multiprocessor or multicore computers, of direct or fifo-based message passing in distributed networks, and of fifo- or interrupt-based event handling in web-based human-machine interaction or sensor handling. Synchronous or quasi-synchronous concurrency is the basis of signal processing, of real-time control, and of safety-critical information acquisition and display. Interfacing existing devices based on these different concurrency principles within *Hop* or other diffuse programming languages will require better understanding of the underlying concurrency models and of the way they can nicely cooperate, a currently ill-resolved problem.

3.2. Web, functional, and reactive programming

We are studying new paradigms for programming Web applications that rely on multi-tier functional programming. We have created a Web programming environment named *Hop*. It relies on a single formalism for programming the server-side and the client-side of the applications as well as for configuring the execution engine.

Hop is a functional language based on the SCHEME programming language. That is, it is a strict functional language, fully polymorphic, supporting side effects, and dynamically type-checked. *Hop* is implemented as an extension of the BIGLOO compiler that we develop. In the past, we have extensively studied static analyses (type systems and inference, abstract interpretations, as well as classical compiler optimizations) to improve the efficiency of compilation in both space and time.

As a *Hop* DSL, we have created *HipHop*, a synchronous orchestration language for web and IoT applications. *HipHop* facilitates the design and programming of complex web/IoT applications by smoothly integrating three computation models and programming styles that have been historically developed in different communities and for different purposes: *i) Transformational programs* that simply compute output values from input values, with comparatively simple interaction with their environment; *ii)* asynchronous concurrent programs that perform interactions between their components or with their environment with uncontrollable timing, using typically network-based communication; and *iii)* synchronous reactive programs that react to external events in a conceptually instantaneous and deterministic way.

3.3. Security of diffuse programs

The main goal of our security research is to provide scalable and rigorous language-based techniques that can be integrated into multi-tier compilers to enforce the security of diffuse programs. Research on language-based security has been carried on before in former Inria teams. In particular previous research has focused on controlling information flow to ensure confidentiality.

Typical language-based solutions to these problems are founded on static analysis, logics, provable cryptography, and compilers that generate correct code by construction. Relying on the multi-tier programming language *Hop* that tames the complexity of writing and analysing secure diffuse applications, we are studying language-based solutions to prominent web security problems such as code injection and cross-site scripting, to name a few.

KERDATA Project-Team

3. Research Program

3.1. Research axis 1: Convergence of HPC and Big Data

The tools and cultures of High Performance Computing and Big Data Analytics have evolved in divergent ways. This is to the detriment of both. However, big computations still generate and are needed to analyze Big Data. As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two ecosystems is crucial to the future.

Our objective is premised on the idea that we must explore the ways in which the major challenges associated with Big Data analytics intersect with, impact, and potentially change the directions now in progress for achieving Exascale computing.

In particular, a key milestone will be to achieve convergence through common abstractions and techniques for data storage and processing in support of complex workflows combining simulations and analytics. Such application workflows will need such a convergence to run on hybrid infrastructures combining HPC systems and clouds (potentially in extension to edge devices, in a complete digital continuum).

Collaboration. *This axis is addressed in close collaboration with María Pérez (UPM), Rob Ross (ANL), Toni Cortes (BSC), Several groups at Argonne National Laboratory and NCSA (Franck Cappello, Rob Ross, Bill Kramer, Tom Peterka).*

Relevant groups with similar interests are the following ones.

- *The group of Jack Dongarra, Innovative Computing Laboratory at University of Tennessee, who is leading international efforts for the convergence of Exascale Computing and Big Data.*
- *The group of Satoshi Matsuoka, RIKEN, working on system software for clouds and HPC.*
- *The group of Ian Foster, Argonne National Laboratory, working on on-demand data analytics and storage for extreme-scale simulations and experiments.*

3.1.1. High-performance storage for concurrent Big Data applications

Storage is a plausible pathway to convergence. In this context, we plan to focus on the needs of concurrent Big Data applications that require high-performance storage, as well as transaction support. Although blobs (binary large objects) are an increasingly popular storage model for such applications, state-of-the-art blob storage systems offer no transaction semantics. This demands users to coordinate data access carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior.

There is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications. In this context, one idea on which we plan to focus our efforts is exploring how blob storage systems could provide built-in, multiblob transactions, while retaining sequential consistency and high throughput under heavy access concurrency.

The early principles of this research direction have already raised interest from our partners at ANL (Rob Ross) and UPM (María Pérez) for potential collaborations. In this direction, the acceptance of our paper on the Týr transactional blob storage system as a Best Student Paper Award Finalist at the SC16 conference [10] is a very encouraging step.

3.1.2. Towards unified data processing techniques for Extreme Computing and Big Data applications

In the high-performance computing area (HPC), the need to get fast and relevant insights from massive amounts of data generated by extreme-scale computations led to the emergence of *in situ processing*. It allows data to be visualized and processed in real-time on the supercomputer generating them, in an interactive way, as they are produced, as opposed to the traditional approach consisting of transferring data off-site after the end of the computation, for offline analysis. As such processing runs on the same resources executing the simulation, if it consumes too many resources, there is a risk to "disturb" the simulation.

Consequently, an alternative approach was proposed (*in transit processing*), as a means to reduce this impact: data are are transferred to some temporary processing resources (with high memory and processing capacities). After this real-time processing, they are moved to persistent storage.

In the Big Data area, the search for real-time, fast analysis was materialized through a different approach: stream-based processing. Such an approach is based on a different abstraction for data, that are seen as a dynamic flow of items to be processed. Stream-based processing and in situ/in transit processing have been developed separately and implemented in different tools in the BDA and HPC areas respectively.

A major challenge from the perspective of the HPC-BDA convergence is their joint use in a unified data processing architecture. This is one of the future research challenges that I plan to address in the near future, by combining ongoing approaches currently active in my team: Damaris and KerA. We started preliminary work within the *"Frameworks"* work package of the HPC-Big Data IPL. Further exploring this convergence is a core direction of our current efforts to build collaborative European projects.

3.2. Research axis 2: Cloud and Edge processing

The recent evolutions in the area of Big Data processing have pointed out some limitations of the initial Map-Reduce model. It is well suited for batch data processing, but less suited for real-time processing of dynamic data streams. New types of data-intensive applications emerge, e.g., for enterprises who need to perform analysis on their stream data in ways that can give fast results (i.e., in real time) at scale (e.g., click-stream analysis and network-monitoring log analysis). Similarly, scientists require fast and accurate data processing techniques in order to analyze their experimental data correctly at scale (e.g., collectively analysis of large data sets distributed in multiple geographically distributed locations).

Our plan is to revisit current data storage and processing techniques to cope with the volatile requirements of data-intensive applications on large-scale dynamic clouds in a cost-efficient way, with a particular focus on streaming. More recently, the strong emergence of edge/fog-based infrastructures leads to to additional challenges for new scenarios involving hybrid cloud/fog/edge systems.

Collaboration. *This axis is addressed in close collaboration with María Pérez (UPM), Kate Keahey. (ANL)*

Relevant groups with similar interests include the following ones.

- *The group of Geoffrey Fox, Indiana University, working on data analytics, cloud data processing, stream processing.*
- *The group at RISE Lab, UC Berkeley, working on real-time stream-based processing and analytics.*
- *The group of Ewa Deelman, USC Information Sciences Institute, working on resource management for workflows in clouds.*

3.2.1. Stream-oriented, Big Data processing on clouds

The state-of-the-art Hadoop Map-Reduce framework cannot deal with stream data applications, as it requires the data to be initially stored in a distributed file system in order to process them. To better cope with the above-mentioned requirements, several systems have been introduced for stream data processing such as Flink [27], Spark [32], Storm [33], and Google MillWheel [34]. These systems keep computation in memory to decrease

latency, and preserve scalability by using data-partitioning or dividing the streams into a set of deterministic batch computations.

However, they are designed to work in dedicated environments and they do not consider the performance variability (i.e., network, I/O, etc.) caused by resource contention in the cloud. This variability may in turn cause high and unpredictable latency when output streams are transmitted to further analysis. Moreover, they overlook the dynamic nature of data streams and the volatility in their computation requirements. Finally, they still address failures in a best-effort manner.

Our objective is to investigate new approaches for reliable, stream Big Data processing on clouds.

3.2.2. Efficient Edge, Cloud and hybrid Edge/Cloud data processing

Today, we are approaching an important technological milestone: applications are generating huge amounts of data and are demanding low-latency responses to their requests. Mobile computing and Internet of Things (IoT) applications are good illustrations of such scenarios. Using only Cloud computing for such scenarios is challenging. Firstly, Cloud resources are most of the time accessed through Internet, hence, data are sent across high-latency wide area networks, which may degrade the performance of applications. Secondly, it may be impossible to send data to the Cloud due to data regulations, national security laws or simply because an Internet connection is not available. Finally, data transmission costs (e.g., Cloud provider fees, carrier costs) could make a business solution impractical.

Edge computing is a new paradigm which aims to address some of these issues. The key idea is to leverage computing and storage resources at the "edge" of the network, i.e., on processing units located close to the data sources. This allows applications to outsource task execution from the main (Cloud) processing data centers to the edge. The development of Edge computing was accelerated by the recent emergence of stream processing, a new model for handling continuous flows of data in real-time, as opposed to batch processing, which typically processes bounded datasets offline.

However, Edge computing is not a silver bullet. Besides being a new concept not fully established in the community, issues like node volatility, limited processing power, high latency between nodes, fault tolerance and data degradation may impact applications depending on the characteristics of the infrastructure.

Some relevant research questions are: How much can one improve (or degrade) the performance of an application by performing data processing closer to the data sources rather than performing it in the cloud? How to progress towards a seamless scheduling and execution of a data analytics workflow and break the limitation the current dual approaches used in preliminary efforts in this area, that rely on manual and empirical deployment of the corresponding dataflow operator graphs, using separate analytics engines for centralized clouds and for edge systems respectively?

Our objective is to try to answer precisely such questions. We are interested in understanding the conditions that enable the usage of Edge or Cloud computing to reduce the time to results and the associated costs. While some state-of-the-art approaches advocate either "100% Cloud" or "100% Edge" solutions, the relative efficiency of a method over the other may vary. Intuitively, it depends on many parameters, including network technology, hardware characteristics, volume of data or computing power, processing framework configuration and application requirements, to cite a few. We plan to study their impact on the overall application performance.

3.3. Research axis 3: Supporting AI across the digital continuum

Integrating and processing high-frequency data streams from multiple sensors scattered over a large territory in a timely manner requires high-performance computing techniques and equipments. For instance, a machine learning earthquake detection solution has to be designed jointly with experts in distributed computing and cyber-infrastructure to enable real-time alerts. Because of the large number of sensors and their high sampling rate, a traditional centralized approach which transfers all data to a single point may be impractical. Our goal is to investigate innovative solutions for the design of efficient data processing infrastructures for a distributed machine learning-based approach.

In particular, building on our previous results in the area of efficient stream processing systems, we aim to explore approaches for unified data storage, processing and machine-learning based analytics across the whole digital continuum (i.e., for highly distributed applications deployed on hybrid edge/cloud/HPC infrastructures). Our ZettaFlow project is targeting a startup creation precisely this area.

Collaboration. *This recently started axis is worked out in close collaboration with the group of Manish Parashar, Rutgers University, and with the LACODAM team at Inria, focused on large-scale collaborative data mining.*

MARACAS Team

3. Research Program

3.1. General description

As presented in the first section, *Computing Networks* is a concept generalizing the study of multi-user systems under the communication perspective. This problematic is partly addressed in the aforementioned references. Optimizing *Computing Networks* relies on exploiting simultaneously multi-user communication capabilities, in the one hand, and storage and computing resources in the other hand. Such optimization needs to cope with various constraints such as energy efficiency or energy harvesting, delays, reliability or network load.

The notion of reliability (used in MARACAS acronym) is central when considered in the most general sense : ultimately, the reliability of a *Computing Network* measures its capability to perform its intended role under some confidence interval. Figure 1 represents the most important performance criteria to be considered to achieve reliable communications. These metrics fit with those considered in 5G and beyond technologies [63].

On the theoretical side, multi-user information theory is a keystone element. It is worth noting that classical information theory focuses on the power-bandwith tradeoff usually referred as Energy Efficiency-Spectral Efficiency (EE-SE) tradeoff (green arrow on 1). However, the other constraints can be efficiently introduced by using a non-asymptotic formulation of the fundamental limits [62], [64] and in association with other tools devoted to the analysis of random processes (queuing theory, ...).

Maracas aims at studying *Computing Networks* from a communication point of view, using the foundations of information theory in association with other theoretical tools related to estimation theory and probability theory.

In particular, Maracas combines techniques from communication and information theory with statistical signal processing, control theory, and game theory. Wireless networks is the emblematic application for Maracas, but other scenarios are appealing for us, such as molecular communications, smart grids or smart buildings.

Several teams at Inria are addressing computing networks, but working on this problem with an emphasis on communication aspects is unique within Inria.

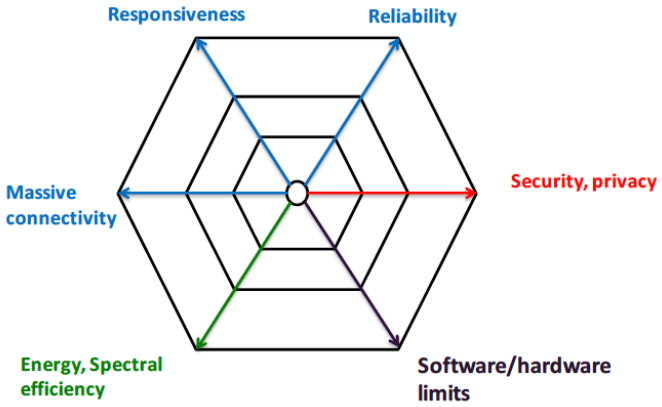


Figure 1. Main metrics for future networks (5G and beyond)

The complexity of *Computing Networks* comes first from the high dimensionality of the problem: i) thousands of nodes, each with up to tens setting parameters and ii) tens variable objective functions to be minimized/maximized.

In addition, the necessary decentralization of the decision process, the non stationary behavior of the network itself (mobility, ON/OFF Switching) and of the data flows, and the necessary reduction of costly feedback and signaling (channel estimation, topology discovering, medium access policies...) are additional features that increase the problem complexity.

The original positioning of Maracas holds in his capability to address three complementary challenges :

1. **to develop a sound mathematical framework inspired by information theory.**
2. **to design algorithms, achieving performance close to these limits.**
3. **to test and validate these algorithms on experimental testbeds.**

3.2. Research program

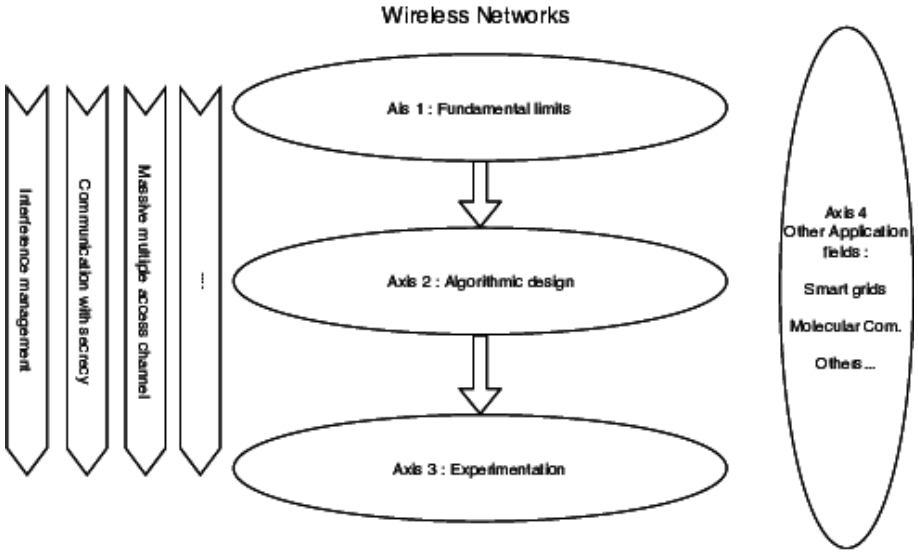


Figure 2. Maracas organization

Our research is organized in 4 research axes:

- **Axis 1 - Fundamental Limits of Reliable Communication Systems**: Information theory is revisited to integrate reliability in the wide sense. The non-asymptotic theory which made progress recently and attracted a lot of interest in the information theory community is a good starting point. But for addressing computing network in a wide sense, it is necessary to go back to the foundation of communication theory and to derive new results, e.g. for non Gaussian channels [8] of for multi-constrained systems [17].

  This also means revisiting the fundamental estimation-detection problem [65] in a general multi-criteria, multi-user framework to derive tractable and meaningful bounds.

As mentioned in the introduction, *Computing Networks* also relies on a data-centric vision, where transmission, storage and processing are jointly optimized. The strategy of *caching at the edge* [57] proposed for cellular networks shows the high potential of considering simultaneously data and network properties. Maracas is willing to extend his skills on source coding aspects to tackle with a data-oriented modeling of *Computing Networks*.

- **Axis 2 - Algorithms and protocols**: Our second objective is to elaborate new algorithms and protocols able to achieve or at least to approach the aforementioned fundamental limits. While the exploration of fundamental limits is helpful to determine the most promising strategies (e.g. relaying, cooperation, interference alignment) to increase system performance, the transformation of these degrees of freedom into real protocols is a non trivial issue. One reason is the exponentially growing complexity of multi-user communication strategies, with the number of users, due to the necessity of some coordination, feedback and signaling. The general problem is a decentralized and dynamic multi-agents multi-criteria optimization problem and the general formulation is a non-linear and non-convex large scale problem.

  The conventional research direction aims at reducing the complexity by relaxing some constraints or by reducing the number of degrees of freedom. For instance, topology interference management is a seducing model used to reduce feedback needs in decentralized wireless networks leading to original and efficient algorithms [67], [59].

  Another emerging research direction relies on using machine learning techniques [54] as a natural evolution of cognitive radio based approaches. Machine learning in the wide sense is not new in radio networks, but the most important works in the past were devoted to reinforcement learning approaches. The use of deep learning (DL) is much more recent, with two important issues : i) identifying the right problems that really need DL algorithms and ii) providing extensive data sets from simulation and real experiments. Our group started to work on this topic in association with Nokia in the joint research lab. As we are not currently expert in deep learning, our primary objective is to identify the strategic problems and to collaborate in the future with Inria experts in DL, and in the long term to contribute not only to the application of these techniques, but also to improve their design according to the constraints of computing networks.
- **Axis 3 - Experimental validation** : With the rapid evolution of network technologies, and their increasing complexity, experimental validation is necessary for two reasons: to get data, and to validate new algorithms on real systems.

  Maracas activity leverages on the FIT/CorteXlab platform (http://www.cortexlab.fr/), and our strong partnerships with leading industry including Nokia Bell Labs, Orange labs, Sigfox or Sequans. Beyond the platform itself which offers a worldwide unique and remotely accessible testbed , Maracas also develops original experimentations exploiting the reproducibility, the remote accessibility, and the deployment facilities to produce original results at the interface of academic and industrial research [1], [10]. FIT/CorteXlab uses the GNU Radio environment to evaluate new multi-user communication systems.

  Our experimental work is developed in collaboration with other Inria teams especially in the Rhone-Alpes centre but also in the context of the future SILECS project https://www.silecs.net/ which will implement the convergence between FIT and Grid'5000 infrastructures in France, in cooperation with European partners and infrastructures. SILECS is a unique framework which will allow us to test our algorithms, to generate data, as required to develop a data-centric approach for computing networks.

  Last but not least, software radio technologies are leaving the confidentiality of research laboratories and are made available to a wide public market with cheap (few euros) programmable equipment, allowing to setup non standard radio systems. The existence of home-made and non official radio systems with legacy ones could prejudice the deployment of Internet of things. Developing efficient algorithms able to detect, analyse and control the spectrum usage is an important issue. Our research on FIT/CorteXlab will contribute to this know-how.

- **Axis 4 - Other application fields** : Even if the wireless network context is still challenging and provides interesting problems, Maracas targets to broaden its exploratory playground from an application perspective. We are looking for new communication systems, or simply other multi-user decentralized systems, for which the theory developed in the context of wireless networks can be useful. Basically, Maracas might address any problem where multi-agents are trying to optimize their common behavior and where the communication performance is critical (e.g. vehicular communications, multi-robots systems, cyberphysical systems). Following this objective, we already studied the problem of missing data recovery in smart grids [11] and the original paradigm of molecular communications [6].

  Of course, the objective of this axis is not to address random topics but to exploit our scientific background on new problems, in collaboration with other academic teams or industry. This is a winning strategy to develop new partnerships, in collaboration with other Inria teams.

MIMOVE Project-Team

3. Research Program

3.1. Introduction

The research questions identified above call for radically new ways in conceiving, developing and running mobile distributed systems. In response to this challenge, MiMove's research aims at enabling next-generation mobile distributed systems that are the focus of the following research topics.

3.2. Emergent mobile distributed systems

Uncertainty in the execution environment calls for designing mobile distributed systems that are able to run in a beforehand unknown, ever-changing context. Nevertheless, the complexity of such change cannot be tackled at system design-time. Emergent mobile distributed systems are systems which, due to their automated, dynamic, environment-dependent composition and execution, *emerge* in a possibly non-anticipated way and manifest *emergent properties*, i.e., both systems and their properties take their complete form only at runtime and may evolve afterwards. This contrasts with the typical software engineering process, where a system is finalized during its design phase. MiMove's research focuses on enabling the emergence of mobile distributed systems while assuring that their required properties are met. This objective builds upon pioneering research effort in the area of *emergent middleware* initiated by members of the team and collaborators [1], [3].

3.3. Large-scale mobile sensing and actuation

The extremely large scale and dynamicity expected in future mobile sensing and actuation systems lead to the clear need for algorithms and protocols for addressing the resulting challenges. More specifically, since connected devices will have the capability to sense physical phenomena, perform computations to arrive at decisions based on the sensed data, and drive actuation to change the environment, enabling proper coordination among them will be key to unlocking their true potential. Although similar challenges have been addressed in the domain of networked sensing, including by members of the team [7], the specific challenges arising from the *extremely large scale* of mobile devices – a great number of which will be attached to people, with uncontrolled mobility behavior – are expected to require a significant rethink in this domain. MiMove's research investigates techniques for efficient coordination of future mobile sensing and actuation systems with a special focus on their dependability.

3.4. Mobile social crowd-sensing

While mobile social sensing opens up the ability of sensing phenomena that may be costly or impossible to sense using embedded sensors (e.g., subjective crowdedness causing discomfort or joyfulness, as in a bus or in a concert) and leading to a feeling of being more socially involved for the citizens, there are unique consequent challenges. Specifically, MiMove's research focuses on the problems involved in the combination of the physically sensed data, which are quantitative and objective, with the mostly qualitative and subjective data arising from social sensing. Enabling the latter calls for introducing mechanisms for incentivising user participation and ensuring the privacy of user data, as well as running empirical studies for understanding the complex social behaviors involved. These objectives build upon previous research work by members of the team on mobile social ecosystems and privacy, as well as a number of efforts and collaborations in the domain of smart cities and transport that have resulted in novel mobile applications enabling empirical studies of social sensing systems.

3.5. Active and passive probing methods

We are developing methods that actively introduce probes in the network to discover properties of the connected devices and network segments. We are focusing in particular on methods to discover properties of home networks (connected devices and their types) and to distinguish if performance bottlenecks lie within the home network versus in the different network segments outside (e.g., Internet access provider, interconnects, or content provider). Our goal is to develop adaptive methods that can leverage the collaboration of the set of available devices (including end-user devices and the home router, depending on which devices are running the measurement software).

We are also developing passive methods that simply observe network traffic to infer the performance of networked applications and the location of performance bottlenecks, as well as to extract patterns of web content consumption. We are working on techniques to collect network traffic both at user's end-devices and at home routers. We also have access to network traffic traces collected on a campus network and on a large European broadband access provider.

3.6. Inferring user online experience

We are developing hybrid measurement methods that combine passive network measurement techniques to infer application performance with techniques from HCI to measure user perception as well as methods to directly measure application quality. We later use the resulting datasets to build models of user perception of network performance based only on data that we can obtain automatically from the user device or from user's traffic observed in the network.

3.7. Real time data analytics

The challenge of deriving insights from the Internet of Things (IoT) has been recognized as one of the most exciting and key opportunities for both academia and industry. The time value of data is crucial for many IoT-based systems requiring *real-time* (or near real-time) *control* and *automation*. Such systems typically collect data continuously produced by "things" (i.e., devices), and analyze them in (sub-) seconds in order to act promptly, e.g., for detecting security breaches of digital systems, for spotting malfunctions of physical assets, for recommending goods and services based on the proximity of potential clients, etc. Hence, they require to both *ingest* and *analyze in real-time* data arriving with different velocity from various IoT data streams.

Existing incremental (online or streaming) techniques for descriptive statistics (e.g., frequency distributions, frequent patterns, etc.) or predictive statistics (e.g., classification, regression) usually assume a good enough quality dataset for mining patterns or training models. However, IoT raw data produced in the wild by sensors embedded in the environment or wearable by users are prone to errors and noise. Effective and efficient algorithms are needed for *detecting* and *repairing data impurities* (for controlling data quality) as well as *understanding data dynamics* (for defining alerts) in real-time, for collections of IoT data streams that might be geographically distributed. Moreover, supervised deep learning and data analytics techniques are challenged by the presence of sparse ground truth data in real IoT applications. Lightweight and adaptive semi-supervised or unsupervised techniques are needed to power real-time anomaly and novelty detection in IoT data streams. The effectiveness of these techniques should be able to reach a useful level through training on a relatively small amount of (preferably unlabeled) data while they can cope distributional characteristics of data evolving over time.

Myriads Project-Team

3. Research Program

3.1. Introduction

In this section, we present our research challenges along four work directions: resource and application management in distributed cloud and fog computing architectures for scaling clouds in Section 3.2 , energy management strategies for greening clouds in Section 3.3 , security and data protection aspects for securing cloud-based information systems and applications in Section 3.4 , and methods for experimenting with clouds in Section 3.5 .

3.2. Scaling fogs and clouds

3.2.1. Resource management in hierarchical clouds

The next generation of utility computing appears to be an evolution from highly centralized clouds towards more decentralized platforms. Today, cloud computing platforms mostly rely on large data centers servicing a multitude of clients from the edge of the Internet. Servicing cloud clients in this manner suggests that locality patterns are ignored: wherever the client issues his/her request from, the request will have to go through the backbone of the Internet provider to the other side of the network where the data center relies. Besides this extra network traffic and this latency overhead that could be avoided, other common centralization drawbacks in this context stand in limitations in terms of security/legal issues and resilience.

At the same time, it appears that network backbones are over-provisioned for most of their usage. This advocates for placing computing resources directly within the backbone network. The general challenge of resource management for such clouds stands in trying to be locality-aware: for the needs of an application, several virtual machines may exchange data. Placing them *close* to each others can significantly improve the performance of the application they compose. More generally, building an overlay network which takes the hierarchical aspects of the platform without being a hierarchical overlay – which comes with load balancing and resilience issues is a challenge by itself.

We expect to integrate the results of these works in the Discovery initiative [33] which aims at revisiting OpenStack to offer a cloud stack able to manage utility computing platforms where computing resources are located in small computing centers in the backbone's PoPs (Point of Presence) and interconnected through the backbone's internal links.

3.2.2. Resource management in fog computing architectures

Fog computing infrastructures are composed of compute, storage and networking resources located at the edge of wide-area networks, in immediate proximity to the end users. Instead of treating the mobile operator's network as a high-latency dumb pipe between the end users and the external service providers, fog platforms aim at deploying cloud functionalities *within* the mobile phone network, inside or close to the mobile access points. Doing so is expected to deliver added value to the content providers and the end users by enabling new types of applications ranging from Internet-of-Things applications to extremely interactive systems (e.g., augmented reality). Simultaneously, it will generate extra revenue streams for the mobile network operators, by allowing them to position themselves as cloud computing operators and to rent their already-deployed infrastructure to content and application providers.

Fog computing platforms have very different geographical distribution compared to traditional clouds. While traditional clouds are composed of many reliable and powerful machines located in a very small number of data centers and interconnected by very high-speed networks, mobile edge cloud are composed of a very large number of points-of-presence with a couple of weak and potentially unreliable servers, interconnected with each other by commodity long-distance networks. This creates new demands for the organization of a scalable mobile edge computing infrastructure, and opens new directions for research.

The main challenges that we plan to address are:

- How should an edge cloud infrastructure be designed such that it remains scalable, fault-tolerant, controllable, energy-efficient, etc.?
- How should applications making use of edge clouds be organized? One promising direction is to explore the extent to which stream-data processing platforms such as Apache Spark and Apache Flink can be adapted to become one of the main application programming paradigms in such environments.

3.2.3. Self-optimizing applications in multi-cloud environments

As the use of cloud computing becomes pervasive, the ability to deploy an application on a multi-cloud infrastructure becomes increasingly important. Potential benefits include avoiding dependence on a single vendor, taking advantage of lower resource prices or resource proximity, and enhancing application availability. Supporting multi-cloud application management involves two tasks. First, it involves selecting an initial multi-cloud application deployment that best satisfies application objectives and optimizes performance and cost. Second, it involves dynamically adapting the application deployment in order to react to changes in execution conditions, application objectives, cloud provider offerings, or resource prices. Handling price changes in particular is becoming increasingly complex. The reason is the growing trend of providers offering sophisticated, dynamic pricing models that allow buying and selling resources of finer granularities for shorter time durations with varying prices.

Although multi-cloud platforms are starting to emerge, these platforms impose a considerable amount of effort on developers and operations engineers, provide no support for dynamic pricing, and lack the responsiveness and scalability necessary for handling highly-distributed, dynamic applications with strict quality requirements. The goal of this work is to develop techniques and mechanisms for automating application management, enabling applications to cope with and take advantage of the dynamic, diverse, multi-cloud environment in which they operate.

The main challenges arising in this context are:

- selecting effective decision-making approaches for application adaptation,
- supporting scalable monitoring and adaptation across multiple clouds,
- performing adaptation actions in a cost-efficient and safe manner.

3.3. Greening clouds

The ICT (Information and Communications Technologies) ecosystem now approaches 5% of world electricity consumption and this ICT energy use will continue to grow fast because of the information appetite of Big Data, large networks and large infrastructures as Clouds that unavoidably leads to large power.

3.3.1. Smart grids and clouds

We propose exploiting Smart Grid technologies to come to the rescue of energy-hungry Clouds. Unlike in traditional electrical distribution networks, where power can only be moved and scheduled in very limited ways, Smart Grids dynamically and effectively adapt supply to demand and limit electricity losses (currently 10% of produced energy is lost during transmission and distribution).

For instance, when a user submits a Cloud request (such as a Google search for instance), it is routed to a data center that processes it, computes the answer and sends it back to the user. Google owns several data centers spread across the world and for performance reasons, the center answering the user's request is more likely to be the one closest to the user. However, this data center may be less energy efficient. This request may have consumed less energy, or a different kind of energy (renewable or not), if it had been sent to this further data center. In this case, the response time would have been increased but maybe not noticeably: a different trade-off between quality of service (QoS) and energy-efficiency could have been adopted.

While Clouds come naturally to the rescue of Smart Grids for dealing with this big data issue, little attention has been paid to the benefits that Smart Grids could bring to distributed Clouds. To our knowledge, no previous work has exploited the Smart Grids potential to obtain and control the energy consumption of entire Cloud infrastructures from underlying facilities such as air conditioning equipment (which accounts for 30% to 50% of a data center's electricity bill) to network resources (which are often operated by several actors) and to computing resources (with their heterogeneity and distribution across multiple data centers). We aim at taking advantage of the opportunity brought by the Smart Grids to exploit renewable energy availability and to optimize energy management in distributed Clouds.

3.3.2. Energy cost models

Cloud computing allows users to outsource the computer resources required for their applications instead of using a local installation. It offers on-demand access to the resources through the Internet with a pay-as-you-go pricing model.However, this model hides the electricity cost of running these infrastructures.

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructures). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users.

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models will be based on experimental measurement campaigns on heterogeneous devices. Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale.

3.3.3. Energy-aware users

In a moderately loaded Cloud, some servers may be turned off when not used for energy saving purpose. Cloud providers can apply resource management strategies to favor idle servers. Some of the existing solutions propose mechanisms to optimize VM scheduling in the Cloud. A common solution is to consolidate the mapping of the VMs in the Cloud by grouping them in a fewer number of servers. The unused servers can then be turned off in order to lower the global electricity consumption.

Indeed, current work focuses on possible levers at the virtual machine suppliers and/or services. However, users are not involved in the choice of using these levers while significant energy savings could be achieved with their help. For example, they might agree to delay slightly the calculation of the response to their applications on the Cloud or accept that it is supported by a remote data center, to save energy or wait for the availability of renewable energy. The VMs are black boxes from the Cloud provider point of view. So, the user is the only one to know the applications running on her VMs.

We plan to explore possible collaborations between virtual machine suppliers, service providers and users of Clouds in order to provide users with ways of participating in the reduction of the Clouds energy consumption. This work will follow two directions: 1) to investigate compromises between power and performance/service quality that cloud providers can offer to their users and to propose them a variety of options adapted to their workload; and 2) to develop mechanisms for each layer of the Cloud software stack to provide users with a quantification of the energy consumed by each of their options as an incentive to become greener.

3.4. Securing clouds

3.4.1. Security monitoring SLO

While the trend for companies to outsource their information system in clouds is confirmed, the problem of securing an information system becomes more difficult. Indeed, in the case of infrastructure clouds, physical

resources are shared between companies (also called tenants) but each tenant controls only parts of the shared resources, and, thanks to virtualization, the information system can be dynamically and automatically reconfigured with added or removed resources (for example starting or stopping virtual machines), or even moved between physical resources (for example using virtual machine migration). Partial control of shared resources brings new classes of attacks between tenants, and security monitoring mechanisms to detect such attacks are better placed out of the tenant-controlled virtual information systems, that is under control of the cloud provider. Dynamic and automatic reconfigurations of the information system make it unfeasible for a tenant's security administrator to setup the security monitoring components to detect attacks, and thus an automated self-adaptable security monitoring service is required.

Combining the two previous statements, there is a need for a dependable, automatic security monitoring service provided to tenants by the cloud provider. Our goal is to address the following challenges to design such a security monitoring service:

1. to define relevant Service-Level Objectives (SLOs) of a security monitoring service, that can figure in the Service-Level Agreement (SLA) signed between a cloud provider and a tenant;
2. to design heuristics to automatically configure provider-controlled security monitoring software components and devices so that SLOs are reached, even during automatic reconfigurations of tenants' information systems;
3. to design evaluation methods for tenants to check that SLOs are reached.

Moreover in challenges 2 and 3 the following sub-challenges must be addressed:

- although SLAs are bi-lateral contracts between the provider and each tenant, the implementation of the contracts is based on shared resources, and thus we must study methods to combine the SLOs;
- the designed methods should have a minimal impact on performance.

3.4.2. Data protection in Cloud-based IoT services

The Internet of Things is becoming a reality. Individuals have their own swarm of connected devices (e.g. smartphone, wearables, and home connected objects) continually collecting personal data. A novel generation of services is emerging exploiting data streams produced by the devices' sensors. People are deprived of control of their personal data as they don't know precisely what data are collected by service providers operating on Internet (oISP), for which purpose they could be used, for how long they are stored, and to whom they are disclosed. In response to privacy concerns the European Union has introduced, with the Global Data Protection Regulation (GDPR), new rules aimed at enforcing the people's rights to personal data protection. The GDPR also gives strong incentives to oISPs to comply. However, today, oISPs can't make their systems GDPR-compliant since they don't have the required technologies. We argue that a new generation of system is mandatory for enabling oISPs to conform to the GDPR. We plan to to design an open source distributed operating system for native implementation of new GDPR rules and ease the programming of compliant cloud-based IoT services. Among the new rules, transparency, right of erasure, and accountability are the most challenging ones to be implemented in IoT environments but could fundamentally increase people's confidence in oISPs. Deployed on individuals' swarms of devices and oISPs' cloud-hosted servers, it will enforce detailed data protection agreements and accountability of oISPs' data processing activities. Ultimately we will show to what extend the new GDPR rules can be implemented for cloud-based IoT services.

3.5. Experimenting with Clouds

Cloud platforms are challenging to evaluate and study with a sound scientific methodology. As with any distributed platform, it is very difficult to gather a global and precise view of the system state. Experiments are not reproducible by default since these systems are shared between several stakeholders. This is even worsen by the fact that microscopic differences in the experimental conditions can lead to drastic changes since typical Cloud applications continuously adapt their behavior to the system conditions.

3.5.1. Experimentation methodologies for clouds

We propose to combine two complementary experimental approaches: direct execution on testbeds such as Grid'5000, that are eminently convincing but rather labor intensive, and simulations (using *e.g.,* SimGrid) that are much more light-weighted, but requires are careful assessment. One specificity of the Myriads team is that we are working on these experimental methodologies *per se*, raising the standards of *good experiments* in our community.

We plan to make SimGrid widely usable beyond research laboratories, in order to evaluate industrial systems and to teach the future generations of cloud practitioners. This requires to frame the specific concepts of Cloud systems and platforms in actionable interfaces. The challenge is to make the framework both easy to use for simple studies in educational settings while modular and extensible to suit the specific needs of every advanced industrial-class users.

We aim at leveraging the convergence opportunities between methodologies by further bridging simulation and real testbeds. The predictions obtained from the simulator should be validated against some real-world experiments obtained on the target production platform, or on a similar platform. This (in)validation of the predicted results often improves the understanding of the modeled system. On the other side, it may even happen that the measured discrepancies are due to some mis-configuration of the real platform that would have been undetected without this (in)validation study. In that sense, the simulator constitutes a precious tool for the quality assurance of real testbeds such as Grid'5000.

Scientists need more help to make their Cloud experiments fully reproducible, in the sprit of Open Science exemplified by the HAL Open Archive, actively backed by Inria. Users still need practical solutions to archive, share and compare the whole experimental settings, including the raw data production (particularly in the case of real testbeds) and their statistical analysis. This is a long lasting task to which we plan to collaborate through the research communities gathered around the Grid'5000 and SimGrid scientific instruments.

Finally, since correction and performance can constitute contradictory goals, it is particularly important to study them jointly. To that extend, we want to bridge the performance studies, that constitute our main scientific heritage, to correction studies leveraging formal techniques. SimGrid already includes support to exhaustively explore the possible executions. We plan to continue this work to ease the use of the relevant formal methods to the experimenter studying Cloud systems.

3.5.2. Use cases

In system research it is important to work on real-world use cases from which we extract requirements inspiring new research directions and with which we can validate the system services and mechanisms we propose. In the framework of our close collaboration with the Data Science Technology department of the LBNL, we will investigate cloud usage for scientific data management. Next-generation scientific discoveries are at the boundaries of datasets, e.g., across multiple science disciplines, institutions and spatial and temporal scales. Today, data integration processes and methods are largely adhoc or manual. A generalized resource infrastructure that integrates knowledge of the data and the processing tasks being performed by the user in the context of the data and resource lifecycle is needed. Clouds provide an important infrastructure platform that can be leveraged by including knowledge for distributed data integration.

NEO Project-Team

3. Research Program

3.1. Stochastic Operations Research

Stochastic Operations Research is a collection of modeling, optimization and numerical computation techniques, aimed at assessing the behavior of man-made systems driven by random phenomena, and at helping to make decisions in such a context.

The discipline is based on applied probability and focuses on effective computations and algorithms. Its core theory is that of Markov chains over discrete state spaces. This family of stochastic processes has, at the same time, a very large modeling capability and the potential of efficient solutions. By "solution" is meant the calculation of some *performance metric*, usually the distribution of some random variable of interest, or its average, variance, etc. This solution is obtained either through exact "analytic" formulas, or numerically through linear algebra methods. Even when not analytically or numerically tractable, Markovian models are always amenable to "Monte-Carlo" simulations with which the metrics can be statistically measured.

An example of this is the success of classical Queueing Theory, with its numerous analytical formulas. Another important derived theory is that of the Markov Decision Processes, which allows to formalize *optimal* decision problems in a random environment. This theory allows to characterize the optimal decisions, and provides algorithms for calculating them.

Strong trends of Operations Research are: a) an increasing importance of multi-criteria multi-agent optimization, and the correlated introduction of Game Theory in the standard methodology; b) an increasing concern of (deterministic) Operations Research with randomness and risk, and the consequent introduction of topics like Chance Constrained Programming and Stochastic Optimization. Data analysis is also more and more present in Operations Research: techniques from statistics, like filtering and estimation, or Artificial Intelligence like clustering, are coupled with modeling in Machine Learning techniques like Q-Learning.

POLARIS Project-Team

3. Research Program

3.1. Sound and Reproducible Experimental Methodology

Participants: Vincent Danjean, Nicolas Gast, Guillaume Huard, Arnaud Legrand, Patrick Loiseau, Jean-Marc Vincent.

Experiments in large scale distributed systems are costly, difficult to control and therefore difficult to reproduce. Although many of these digital systems have been built by men, they have reached such a complexity level that we are no longer able to study them like artificial systems and have to deal with the same kind of experimental issues as natural sciences. The development of a sound experimental methodology for the evaluation of resource management solutions is among the most important ways to cope with the growing complexity of computing environments. Although computing environments come with their own specific challenges, we believe such general observation problems should be addressed by borrowing good practices and techniques developed in many other domains of science.

This research theme builds on a transverse activity on *Open science and reproducible research* and is organized into the following two directions: (1) *Experimental design* (2) *Smart monitoring and tracing*. As we will explain in more detail hereafter, these transverse activity and research directions span several research areas and our goal within the POLARIS project is foremost to transfer original ideas from other domains of science to the distributed and high performance computing community.

3.2. Multi-Scale Analysis and Visualization

Participants: Vincent Danjean, Guillaume Huard, Arnaud Legrand, Jean-Marc Vincent, Panayotis Mertikopoulos.

As explained in the previous section, the first difficulty encountered when modeling large scale computer systems is to observe these systems and extract information on the behavior of both the architecture, the middleware, the applications, and the users. The second difficulty is to *visualize* and *analyze* such *multi-level traces to understand how the performance* of the application *can be improved*. While a lot of efforts are put into visualizing scientific data, in comparison little effort have gone into to developing techniques specifically tailored for understanding the behavior of distributed systems. Many visualization tools have been developed by renowned HPC groups since decades (e.g., BSC [91], Jülich and TU Dresden [90], [61], UIUC [79], [94], [82] and ANL [107], Inria Bordeaux [67] and Grenoble [109], ...) but most of these tools build on the classical information visualization mantra [99] that consists in always first presenting an overview of the data, possibly by plotting everything if computing power allows, and then to allow users to zoom and filter, providing details on demand. However in our context, the amount of data comprised in such traces is several orders of magnitude larger than the number of pixels on a screen and displaying even a small fraction of the trace leads to harmful visualization artifacts [86]. Such traces are typically made of events that occur at very different time and space scales, which unfortunately hinders classical approaches. Such visualization tools have focused on easing interaction and navigation in the trace (through gantcharts, intuitive filters, pie charts and kiviats) but they are very difficult to maintain and evolve and they require some significant experience to identify performance bottlenecks.

Therefore many groups have more recently proposed in combination to these tools some techniques to help identifying the structure of the application or regions (applicative, spatial or temporal) of interest. For example, researchers from the SDSC [89] propose some segment matching techniques based on clustering (Euclidean or Manhattan distance) of start and end dates of the segments that enables to reduce the amount of information to display. Researchers from the BSC use clustering, linear regression and Kriging techniques [98], [85], [78] to identify and characterize (in term of performance and resource usage) application phases and present aggregated representations of the trace [97]. Researchers from Jülich and TU Darmstadt have proposed techniques to identify specific communication patterns that incur wait states [104], [54]

3.3. Fast and Faithful Performance Prediction of Very Large Systems

Participants: Jonatha Anselmi, Vincent Danjean, Bruno Gaujal, Arnaud Legrand, Florence Perronnin, Jean-Marc Vincent.

Evaluating the scalability, robustness, energy consumption and performance of large infrastructures such as exascale platforms and clouds raises severe methodological challenges. The complexity of such platforms mandates empirical evaluation but direct experimentation via an application deployment on a real-world testbed is often limited by the few platforms available at hand and is even sometimes impossible (cost, access, early stages of the infrastructure design, ...). Unlike direct experimentation via an application deployment on a real-world testbed, simulation enables fully repeatable and configurable experiments that can often be conducted quickly for arbitrary hypothetical scenarios. In spite of these promises, current simulation practice is often not conducive to obtaining scientifically sound results. To date, most simulation results in the parallel and distributed computing literature are obtained with simulators that are ad hoc, unavailable, undocumented, and/or no longer maintained. For instance, Naicken et al. [53] point out that out of 125 recent papers they surveyed that study peer-to-peer systems, 52% use simulation and mention a simulator, but 72% of them use a custom simulator. As a result, most published simulation results build on throw-away (short-lived and non validated) simulators that are specifically designed for a particular study, which prevents other researchers from building upon it. There is thus a strong need for recognized simulation frameworks by which simulation results can be reproduced, further analyzed and improved.

The *SimGrid* simulation toolkit [65], whose development is partially supported by POLARIS, is specifically designed for studying large scale distributed computing systems. It has already been successfully used for simulation of grid, volunteer computing, HPC, cloud infrastructures and we have constantly invested on the software quality, the scalability [57] and the validity of the underlying network models [55], [102]. Many simulators of MPI applications have been developed by renowned HPC groups (e.g., at SDSC [100], BSC [51], UIUC [108], Sandia Nat. Lab. [103], ORNL [64] or ETH Zürich [80] for the most prominent ones). Yet, to scale most of them build on restrictive network and application modeling assumptions that make them difficult to extend to more complex architectures and to applications that do not solely build on the MPI API. Furthermore, simplistic modeling assumptions generally prevent to faithfully predict execution times, which limits the use of simulation to indication of gross trends at best. Our goal is to improve the quality of SimGrid to the point where it can be used effectively on a daily basis by practitioners to *reproduce the dynamic of real HPC systems*.

We also develop another simulation software, *PSI* (Perfect SImulator) [69], [62], dedicated to the simulation of very large systems that can be modeled as Markov chains. PSI provides a set of simulation kernels for Markov chains specified by events. It allows one to sample stationary distributions through the Perfect Sampling method (pioneered by Propp and Wilson [92]) or simply to generate trajectories with a forward Monte-Carlo simulation leveraging time parallel simulation (pioneered by Fujimoto [73], Lin and Lazowska [84]). One of the strength of the PSI framework is its expressiveness that allows us to easily study networks with finite and infinite capacity queues [63]. Although PSI already allows to simulate very large and complex systems, our main objective is to push its scalability even further and *improve its capabilities by one or several orders of magnitude*.

3.4. Local Interactions and Transient Analysis in Adaptive Dynamic Systems

Participants: Jonatha Anselmi, Nicolas Gast, Bruno Gaujal, Florence Perronnin, Jean-Marc Vincent, Panayotis Mertikopoulos.

Many systems can be effectively described by stochastic population models. These systems are composed of a set of n entities interacting together and the resulting stochastic process can be seen as a continuous-time Markov chain with a finite state space. Many numerical techniques exist to study the behavior of Markov chains, to solve stochastic optimal control problems [93] or to perform model-checking [52]. These techniques, however, are limited in their applicability, as they suffer from the *curse of dimensionality*: the state-space grows exponentially with n.

This results in the need for approximation techniques. Mean field analysis offers a viable, and often very accurate, solution for large n. The basic idea of the mean field approximation is to count the number of entities that are in a given state. Hence, the fluctuations due to stochasticity become negligible as the number of entities grows. For large n, the system becomes essentially deterministic. This approximation has been originally developed in statistical mechanics for vary large systems composed of more than 10^{20} particles (called entities here). More recently, it has been claimed that, under some conditions, this approximation can be successfully used for stochastic systems composed of a few tens of entities. The claim is supported by various convergence results [74], [83], [106], and has been successfully applied in various domains: wireless networks [56], computer-based systems [77], [88], [101], epidemic or rumour propagation [66], [81] and bike-sharing systems [70]. It is also used to develop distributed control strategies [105], [87] or to construct approximate solutions of stochastic model checking problems [58], [59], [60].

Within the POLARIS project, we will continue developing both the theory behind these approximation techniques and their applications. Typically, these techniques require a homogeneous population of objects where the dynamics of the entities depend only on their state (the state space of each object must not scale with n the number of objects) but neither on their identity nor on their spatial location. Continuing our work in [74], we would like to be able to handle heterogeneous or uncertain dynamics. Typical applications are caching mechanisms [77] or bike-sharing systems [71]. A second point of interest is the use of mean field or large deviation asymptotics to compute the time between two regimes [96] or to reach an equilibrium state. Last, mean-field methods are mostly descriptive and are used to analyse the performance of a given system. We wish to extend their use to solve optimal control problems. In particular, we would like to implement numerical algorithms that use the framework that we developed in [75] to build distributed control algorithms [68] and optimal pricing mechanisms [76].

3.5. Distributed Learning in Games and Online Optimization

Participants: Nicolas Gast, Bruno Gaujal, Arnaud Legrand, Patrick Loiseau, Panayotis Mertikopoulos, Bary Pradelski.

Game theory is a thriving interdisciplinary field that studies the interactions between competing optimizing agents, be they humans, firms, bacteria, or computers. As such, game-theoretic models have met with remarkable success when applied to complex systems consisting of interdependent components with vastly different (and often conflicting) objectives – ranging from latency minimization in packet-switched networks to throughput maximization and power control in mobile wireless networks.

In the context of large-scale, decentralized systems (the core focus of the POLARIS project), it is more relevant to take an inductive, "bottom-up" approach to game theory, because the components of a large system cannot be assumed to perform the numerical calculations required to solve a very-large-scale optimization problem. In view of this, POLARIS' overarching objective in this area is to *develop novel algorithmic frameworks that offer robust performance guarantees when employed by all interacting decision-makers.*

A key challenge here is that most of the literature on learning in games has focused on *static* games with a *finite number of actions* per player [72], [95]. While relatively tractable, such games are ill-suited to practical applications where players pick an action from a continuous space or when their payoff functions evolve over time – this being typically the case in our target applications (e.g., routing in packet-switched networks or energy-efficient throughput maximization in wireless). On the other hand, the framework of online convex optimization typically provides worst-case performance bounds on the learner's *regret* that the agents can attain irrespectively of how their environment varies over time. However, if the agents' environment is determined chiefly by their interactions these bounds are fairly loose, so more sophisticated convergence criteria should be applied.

From an algorithmic standpoint, a further challenge occurs when players can only observe their own payoffs (or a perturbed version thereof). In this bandit-like setting regret-matching or trial-and-error procedures guarantee convergence to an equilibrium in a weak sense in certain classes of games. However, these results apply exclusively to static, finite games: learning in games with continuous action spaces and/or nonlinear

payoff functions cannot be studied within this framework. Furthermore, even in the case of finite games, the complexity of the algorithms described above is not known, so it is impossible to decide a priori which algorithmic scheme can be applied to which application.

RESIST Team

3. Research Program

3.1. Overview

The Resist project aims at designing, implementing and validating novel models, algorithms and tools to **make networked systems elastic and resilient so as to enhance their scalability and security**, assuming users, applications and devices whose volume and heterogeneity will continue to increase.

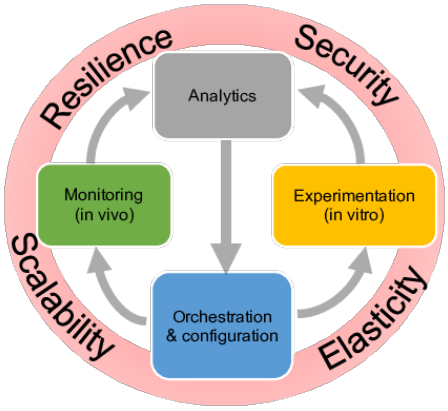


Figure 1. The Resist project

Softwarization of networks and **data analytics** are key enablers to design intelligent methods to orchestrate – *i.e.* configure in a synchronized and distributed manner – both network and system resources. Intelligent **orchestration** leverages relevant data for decision-making using **data analytics**. Input data reflecting the past, current and even future (predicted) states of the system are used to build relevant knowledge. Two approaches are pursued to generate knowledge and to validate orchestration decisions. First, a running system can be **monitored in vivo**. Second, **in vitro experimentation** in a controlled environment (simulators, emulators and experimental platforms) is helpful to reproduce a running system with a high reliability and under different hypotheses. Monitoring and experimentation are steered and configured through orchestration according to the two intertwined loops illustrated in Figure 1 .

Accordingly Resist is thus structured into four main research objectives (activities) namely Monitoring, Experimentation, Analytics and Orchestration.

3.2. Monitoring

The evolving nature of the Internet ecosystem and its continuous growth in size and heterogeneity call for a better understanding of its characteristics, limitations, and dynamics, both locally and globally so as to improve application and protocol design, detect and correct anomalous behaviors, and guarantee performance.

To face these scalability issues, **appropriate monitoring models, methods and algorithms are required for data collection, analysis and sharing** from which knowledge about Internet traffic and usage can be extracted. Measuring and collecting traces necessitate user-centered and data-driven paradigms to cover the wide scope of heterogeneous user activities and perceptions. In this perspective, we propose monitoring algorithms and architectures for large scale environments involving mobile and Internet of Things (IoT) devices.

Resist also assesses **the impact of the Internet infrastructure evolution integrating network softwarization on monitoring**, for example the need for dedicated measurement methodologies. We take into account not only the technological specifics of such paradigms for their monitoring but also the ability to use them for collecting, storing and processing monitoring data in an accurate and cost-effective manner.

Crowd-sourcing and third-party involvement are gaining in popularity, paving the way for massively distributed and collaborative monitoring. We thus investigate opportunistic mobile crowdsensing in order to collect user activity logs along with contextual information (social, demographic, professional) to effectively measure end-users' **Quality of Experience**. However, collaborative monitoring raises serious concerns regarding trust and sensitive data sharing (open data). Data anonymization and sanitization need to be carefully addressed.

3.3. Experimentation

Of paramount importance in our target research context is experimental validation using testbeds, simulators and emulators. In addition to using various existing experimentation methodologies, Resist contributes in **advancing the state of the art in experimentation methods and experimental research practices**, particularly focusing on elasticity and resilience.

We develop and deploy testbeds and emulators for **experimentation with new networking paradigms** such as SDN and NFV, to enable large-scale in-vitro experiments combining all aspects of Software-Defined Infrastructures (server virtualization, SDN/NFV, storage). Such fully controlled environments are particularly suitable for our experiments on resilience, as they ease the management of fault injection features.

We are playing a central role in the development of the Grid'5000 testbed [44] and our objective is to reinforce our collaborations with other testbeds, towards a **testbed federation** in order to enable experiments to scale to multiple testbeds, providing a diverse environment reflecting the Internet itself.

Moreover, our research focuses on extending the infrastructure virtualization capabilities of our Distem [47] emulator, which provides a flexible software-based experimental environment.

Finally, methodological aspects are also important for ensuring **trustworthy and reproducible experiments**, and raises many challenges regarding testbed design, experiment description and orchestration, along with automated or assisted provenance data collection [45].

3.4. Analytics

A large volume of data is processed as part of the operations and management of networked systems. These include traditional monitoring data generated by network components and components' configuration data, but also data generated by dedicated network and system probes.

Understanding and predicting security incidents or system ability to scale requires the elaboration of novel **data analytics techniques** capable to cope with large volumes of data generated from various sources, in various formats, possibly incomplete, non-fully described or even encrypted.

We use machine learning techniques (*e.g.* Topological Data Analysis or multilayer perceptrons) and leverage our domain knowledge to fine-tune them. For instance, machine learning on network data requires the definition of new distance metrics capable to capture the properties of network configurations, packets and flows similarly to edge detection in image processing. Resist contributes to developing and making publicly available an **analytics framework dedicated to networked systems** to support Intelligence-Defined Networked Systems.

Specifically, the goal of the Resist analytics framework is to facilitate the extraction of knowledge useful for **detecting, classifying or predicting security or scalability issues**. The extracted knowledge is then leveraged for orchestration purposes to achieve system elasticity and guarantee its resilience. Indeed, predicting when, where and how issues will occur is very helpful in deciding the provisioning of resources at the right time and place. Resource provisioning can be done either reactively to solve the issues or proactively to prepare the networked system for absorbing the incident (resiliency) in a timely manner thanks to its elasticity.

While the current trend is towards centralization where the collected data is exported to the cloud for processing, we seek to extend this model by also developing and evaluating novel approaches in which **data analytics is seamlessly embedded within the monitored systems**. This combination of big data analytics with network softwarization enablers (SDN, NFV) can enhance the scalability of the monitoring and analytics infrastructure.

3.5. Orchestration

The ongoing transformations in the Internet ecosystem including network softwarization and cloudification bring new management challenges in terms of service and resource orchestration. Indeed, the growing sophistication of Internet applications and the complexity of services deployed to support them require novel models, architectures and algorithms for their automated **configuration** and **provisioning**. Network applications are more and more instantiated through the **composition of services, including virtualized hardware and software resources**, that are offered by **multiple providers** and are subject to changes and updates over time. In this dynamic context, efficient orchestration becomes fundamental for ensuring performance, resilience and security of such applications. We are investigating the chaining of different functions for supporting the security protection of smart devices, based on the networking behavior of their applications.

From a resilience viewpoint, this orchestration at the network level allows the dynamic **reconfiguration of resources** to absorb the effects of congestions, such as link-flooding behaviors. The goal is to drastically reduce the effects of these congestions by imposing dynamic policies on all traffic where the network will adapt itself until it reaches a stable state. We also explore mechanisms for **detecting and remediating potential dysfunctions** within a virtualized network. Corrective operations can be performed through dynamically composed VNFs (Virtualized Network Functions) based on available resources, their dependencies (horizontal and vertical), and target service constraints. We also conduct research on verification methods for automatically assessing and validating the composed chains.

From a security viewpoint, this orchestration provides **prevention mechanisms** that capture adversaries' intentions early and **enforces security policies** in advance through the available resources, to be able to proactively mitigate their attacks. We mainly rely on the results obtained in our research activity on security analytics to build such policies, and the orchestration part focuses on the required algorithms and methods for their automation.

RMOD Project-Team

3. Research Program

3.1. Software Reengineering

Strong coupling among the parts of an application severely hampers its evolution. Therefore, it is crucial to answer the following questions: How to support the substitution of certain parts while limiting the impact on others? How to identify reusable parts? How to modularize an object-oriented application?

Having good classes does not imply a good application layering, absence of cycles between packages and reuse of well-identified parts. Which notion of cohesion makes sense in presence of late-binding and programming frameworks? Indeed, frameworks define a context that can be extended by subclassing or composition: in this case, packages can have a low cohesion without being a problem for evolution. How to obtain algorithms that can be used on real cases? Which criteria should be selected for a given remodularization?

To help us answer these questions, we work on enriching Moose, our reengineering environment, with a new set of analyses [31], [30]. We decompose our approach in three main and potentially overlapping steps:

1. Tools for understanding applications,
2. Remodularization analyses,
3. Software Quality.

3.1.1. Tools for understanding applications

Context and Problems. We are studying the problems raised by the understanding of applications at a larger level of granularity such as packages or modules. We want to develop a set of conceptual tools to support this understanding.

Some approaches based on Formal Concept Analysis (FCA) [59] show that such an analysis can be used to identify modules. However the presented examples are too small and not representative of real code.

Research Agenda.

FCA provides an important approach in software reengineering for software understanding, design anomalies detection and correction, but it suffers from two problems: (i) it produces lattices that must be interpreted by the user according to his/her understanding of the technique and different elements of the graph; and, (ii) the lattice can rapidly become so big that one is overwhelmed by the mass of information and possibilities [20]. We look for solutions to help people putting FCA to real use.

3.1.2. Remodularization analyses

Context and Problems. It is a well-known practice to layer applications with bottom layers being more stable than top layers [47]. Until now, few works have attempted to identify layers in practice: Mudpie [61] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [60], [55] seems to be adapted for such a task but there is no serious empirical experience that validates this claim. From the side of remodularization algorithms, many were defined for procedural languages [43]. However, object-oriented programming languages bring some specific problems linked with late-binding and the fact that a package does not have to be systematically cohesive since it can be an extension of another one [62], [34].

As we are designing and evaluating algorithms and analyses to remodularize applications, we also need a way to understand and assess the results we are obtaining.

Research Agenda. We work on the following items:

Layer identification. We propose an approach to identify layers based on a semi-automatic classification of package and class interrelationships that they contain. However, taking into account the wish or knowledge of the designer or maintainer should be supported.

Cohesion Metric Assessment. We are building a validation framework for cohesion/coupling metrics to determine whether they actually measure what they promise to. We are also compiling a number of traditional metrics for cohesion and coupling quality metrics to evaluate their relevance in a software quality setting.

3.1.3. Software Quality

Research Agenda. Since software quality is fuzzy by definition and a lot of parameters should be taken into account we consider that defining precisely a unique notion of software quality is definitively a Grail in the realm of software engineering. The question is still relevant and important. We work on the two following items:

Quality models. We studied existing quality models and the different options to combine indicators — often, software quality models happily combine metrics, but at the price of losing the explicit relationships between the indicator contributions. There is a need to combine the results of one metric over all the software components of a system, and there is also the need to combine different metric results for any software component. Different combination methods are possible that can give very different results. It is therefore important to understand the characteristics of each method.

Bug prevention. Another aspect of software quality is validating or monitoring the source code to avoid the emergence of well known sources of errors and bugs. We work on how to best identify such common errors, by trying to identify earlier markers of possible errors, or by helping identifying common errors that programmers did in the past.

3.2. Language Constructs for Modular Design

While the previous axis focuses on how to help remodularizing existing software, this second research axis aims at providing new language constructs to build more flexible and recomposable software. We will build on our work on traits [57], [32] and classboxes [21] but also start to work on new areas such as isolation in dynamic languages. We will work on the following points: (1) Traits and (2) Modularization as a support for isolation.

3.2.1. Traits-based program reuse

Context and Problems. Inheritance is well-known and accepted as a mechanism for reuse in object-oriented languages. Unfortunately, due to the coarse granularity of inheritance, it may be difficult to decompose an application into an optimal class hierarchy that maximizes software reuse. Existing schemes based on single inheritance, multiple inheritance, or mixins, all pose numerous problems for reuse.

To overcome these problems, we designed a new composition mechanism called Traits [57], [32]. Traits are pure units of behavior that can be composed to form classes or other traits. The trait composition mechanism is an alternative to multiple or mixin inheritance in which the composer has full control over the trait composition. The result enables more reuse than single inheritance without introducing the drawbacks of multiple or mixin inheritance. Several extensions of the model have been proposed [29], [51], [22], [33] and several type systems were defined [35], [58], [52], [45].

Traits are reusable building blocks that can be explicitly composed to share methods across unrelated class hierarchies. In their original form, traits do not contain state and cannot express visibility control for methods. Two extensions, stateful traits and freezable traits, have been proposed to overcome these limitations. However, these extensions are complex both to use for software developers and to implement for language designers.

Research Agenda: Towards a pure trait language. We plan distinct actions: (1) a large application of traits, (2) assessment of the existing trait models and (3) bootstrapping a pure trait language.

- To evaluate the expressiveness of traits, some hierarchies were refactored, showing code reuse [24]. However, such large refactorings, while valuable, may not exhibit all possible composition problems, since the hierarchies were previously expressed using single inheritance and following certain patterns. We want to redesign from scratch the collection library of Smalltalk (or part of it). Such a redesign should on the one hand demonstrate the added value of traits on a real large and redesigned library and on the other hand foster new ideas for the bootstrapping of a pure trait-based language.

  In particular we want to reconsider the different models proposed (stateless [32], stateful [23], and freezable [33]) and their operators. We will compare these models by (1) implementing a trait-based collection hierarchy, (2) analyzing several existing applications that exhibit the need for traits. Traits may be flattened [50]. This is a fundamental property that confers to traits their simplicity and expressiveness over Eiffel's multiple inheritance. Keeping these aspects is one of our priority in forthcoming enhancements of traits.
- Alternative trait models. This work revisits the problem of adding state and visibility control to traits. Rather than extending the original trait model with additional operations, we use a fundamentally different approach by allowing traits to be lexically nested within other modules. This enables traits to express (shared) state and visibility control by hiding variables or methods in their lexical scope. Although the traits' "flattening property" no longer holds when they can be lexically nested, the combination of traits with lexical nesting results in a simple and more expressive trait model. We formally specify the operational semantics of this combination. Lexically nested traits are fully implemented in AmbientTalk, where they are used among others in the development of a Morphic-like UI framework.
- We want to evaluate how inheritance can be replaced by traits to form a new object model. For this purpose we will design a minimal reflective kernel, inspired first from ObjVlisp [28] then from Smalltalk [38].

3.2.2. Reconciling Dynamic Languages and Isolation

Context and Problems. More and more applications require dynamic behavior such as modification of their own execution (often implemented using reflective features [42]). For example, F-script allows one to script Cocoa Mac-OS X applications and Lua is used in Adobe Photoshop. Now in addition more and more applications are updated on the fly, potentially loading untrusted or broken code, which may be problematic for the system if the application is not properly isolated. Bytecode checking and static code analysis are used to enable isolation, but such approaches do not really work in presence of dynamic languages and reflective features. Therefore there is a tension between the need for flexibility and isolation.

Research Agenda: Isolation in dynamic and reflective languages. To solve this tension, we will work on *Sure*, a language where isolation is provided by construction: as an example, if the language does not offer field access and its reflective facilities are controlled, then the possibility to access and modify private data is controlled. In this context, layering and modularizing the meta-level [25], as well as controlling the access to reflective features [26], [27] are important challenges. We plan to:

- Study the isolation abstractions available in erights (http://www.erights.org) [49], [48], and Java's class loader strategies [44], [39].
- Categorize the different reflective features of languages such as CLOS [41], Python and Smalltalk [53] and identify suitable isolation mechanisms and infrastructure [36].
- Assess different isolation models (access rights, capabilities [54],...) and identify the ones adapted to our context as well as different access and right propagation.
- Define a language based on
  - the decomposition and restructuring of the reflective features [25],

- the use of encapsulation policies as a basis to restrict the interfaces of the controlled objects [56],
- the definition of method modifiers to support controlling encapsulation in the context of dynamic languages.

An open question is whether, instead of providing restricted interfaces, we could use traits to grant additional behavior to specific instances: without trait application, the instances would only exhibit default public behavior, but with additional traits applied, the instances would get extra behavior. We will develop *Sure*, a modular extension of the reflective kernel of Smalltalk (since it is one of the languages offering the largest set of reflective features such as pointer swapping, class changing, class definition,...) [53].

ROMA Project-Team

3. Research Program

3.1. Algorithms for probabilistic environments

There are two main research directions under this research theme. In the first one, we consider the problem of the efficient execution of applications in a failure-prone environment. Here, probability distributions are used to describe the potential behavior of computing platforms, namely when hardware components are subject to faults. In the second research direction, probability distributions are used to describe the characteristics and behavior of applications.

3.1.1. Application resilience

An application is resilient if it can successfully produce a correct result in spite of potential faults in the underlying system. Application resilience can involve a broad range of techniques, including fault prediction, error detection, error containment, error correction, checkpointing, replication, migration, recovery, etc. Faults are quite frequent in the most powerful existing supercomputers. The Jaguar platform, which ranked third in the TOP 500 list in November 2011 [44], had an average of 2.33 faults per day during the period from August 2008 to February 2010 [68]. The mean-time between faults of a platform is inversely proportional to its number of components. Progresses will certainly be made in the coming years with respect to the reliability of individual components. However, designing and building high-reliability hardware components is far more expensive than using lower reliability top-of-the-shelf components. Furthermore, low-power components may not be available with high-reliability. Therefore, it is feared that the progresses in reliability will far from compensate the steady projected increase of the number of components in the largest supercomputers. Already, application failures have a huge computational cost. In 2008, the DARPA white paper on "System resilience at extreme scale" [43] stated that high-end systems wasted 20% of their computing capacity on application failure and recovery.

In such a context, any application using a significant fraction of a supercomputer and running for a significant amount of time will have to use some fault-tolerance solution. It would indeed be unacceptable for an application failure to destroy centuries of CPU-time (some of the simulations run on the Blue Waters platform consumed more than 2,700 years of core computing time [39] and lasted over 60 hours; the most time-consuming simulations of the US Department of Energy (DoE) run for weeks to months on the most powerful existing platforms [42]).

Our research on resilience follows two different directions. On the one hand we design new resilience solutions, either generic fault-tolerance solutions or algorithm-based solutions. On the other hand we model and theoretically analyze the performance of existing and future solutions, in order to tune their usage and help determine which solution to use in which context.

3.1.2. Scheduling strategies for applications with a probabilistic behavior

Static scheduling algorithms are algorithms where all decisions are taken before the start of the application execution. On the contrary, in non-static algorithms, decisions may depend on events that happen during the execution. Static scheduling algorithms are known to be superior to dynamic and system-oriented approaches in stable frameworks [50], [56], [57], [67], that is, when all characteristics of platforms and applications are perfectly known, known a priori, and do not evolve during the application execution. In practice, the prediction of application characteristics may be approximative or completely infeasible. For instance, the amount of computations and of communications required to solve a given problem in parallel may strongly depend on some input data that are hard to analyze (this is for instance the case when solving linear systems using full pivoting).

We plan to consider applications whose characteristics change dynamically and are subject to uncertainties. In order to benefit nonetheless from the power of static approaches, we plan to model application uncertainties and variations through probabilistic models, and to design for these applications scheduling strategies that are either static, or partially static and partially dynamic.

3.2. Platform-aware scheduling strategies

In this theme, we study and design scheduling strategies, focusing either on energy consumption or on memory behavior. In other words, when designing and evaluating these strategies, we do not limit our view to the most classical platform characteristics, that is, the computing speed of cores and accelerators, and the bandwidth of communication links.

In most existing studies, a single optimization objective is considered, and the target is some sort of absolute performance. For instance, most optimization problems aim at the minimization of the overall execution time of the application considered. Such an approach can lead to a very significant waste of resources, because it does not take into account any notion of efficiency nor of yield. For instance, it may not be meaningful to use twice as many resources just to decrease by 10% the execution time. In all our work, we plan to look only for algorithmic solutions that make a "clever" usage of resources. However, looking for the solution that optimizes a metric such as the efficiency, the energy consumption, or the memory-peak minimization, is doomed for the type of applications we consider. Indeed, in most cases, any optimal solution for such a metric is a sequential solution, and sequential solutions have prohibitive execution times. Therefore, it becomes mandatory to consider multi-criteria approaches where one looks for trade-offs between some user-oriented metrics that are typically related to notions of Quality of Service—execution time, response time, stretch, throughput, latency, reliability, etc.—and some system-oriented metrics that guarantee that resources are not wasted. In general, we will not look for the Pareto curve, that is, the set of all dominating solutions for the considered metrics. Instead, we will rather look for solutions that minimize some given objective while satisfying some bounds, or "budgets", on all the other objectives.

3.2.1. Energy-aware algorithms

Energy-aware scheduling has proven an important issue in the past decade, both for economical and environmental reasons. Energy issues are obvious for battery-powered systems. They are now also important for traditional computer systems. Indeed, the design specifications of any new computing platform now always include an upper bound on energy consumption. Furthermore, the energy bill of a supercomputer may represent a significant share of its cost over its lifespan.

Technically, a processor running at speed s dissipates s^α watts per unit of time with $2 \leq \alpha \leq 3$ [48], [49], [54]; hence, it consumes $s^\alpha \times d$ joules when operated during d units of time. Therefore, energy consumption can be reduced by using speed scaling techniques. However it was shown in [69] that reducing the speed of a processor increases the rate of transient faults in the system. The probability of faults increases exponentially, and this probability cannot be neglected in large-scale computing [65]. In order to make up for the loss in *reliability* due to the energy efficiency, different models have been proposed for fault tolerance: (i) *re-execution* consists in re-executing a task that does not meet the reliability constraint [69]; (ii) *replication* consists in executing the same task on several processors simultaneously, in order to meet the reliability constraints [47]; and (iii) *checkpointing* consists in "saving" the work done at some certain instants, hence reducing the amount of work lost when a failure occurs [64].

Energy issues must be taken into account at all levels, including the algorithm-design level. We plan to both evaluate the energy consumption of existing algorithms and to design new algorithms that minimize energy consumption using tools such as resource selection, dynamic frequency and voltage scaling, or powering-down of hardware components.

3.2.2. Memory-aware algorithms

For many years, the bandwidth between memories and processors has increased more slowly than the computing power of processors, and the latency of memory accesses has been improved at an even slower

pace. Therefore, in the time needed for a processor to perform a floating point operation, the amount of data transferred between the memory and the processor has been decreasing with each passing year. The risk is for an application to reach a point where the time needed to solve a problem is no longer dictated by the processor computing power but by the memory characteristics, comparable to the *memory wall* that limits CPU performance. In such a case, processors would be greatly under-utilized, and a large part of the computing power of the platform would be wasted. Moreover, with the advent of multicore processors, the amount of memory per core has started to stagnate, if not to decrease. This is especially harmful to memory intensive applications. The problems related to the sizes and the bandwidths of memories are further exacerbated on modern computing platforms because of their deep and highly heterogeneous hierarchies. Such a hierarchy can extend from core private caches to shared memory within a CPU, to disk storage and even tape-based storage systems, like in the Blue Waters supercomputer [40]. It may also be the case that heterogeneous cores are used (such as hybrid CPU and GPU computing), and that each of them has a limited memory.

Because of these trends, it is becoming more and more important to precisely take memory constraints into account when designing algorithms. One must not only take care of the amount of memory required to run an algorithm, but also of the way this memory is accessed. Indeed, in some cases, rather than to minimize the amount of memory required to solve the given problem, one will have to maximize data reuse and, especially, to minimize the amount of data transferred between the different levels of the memory hierarchy (minimization of the volume of memory inputs-outputs). This is, for instance, the case when a problem cannot be solved by just using the in-core memory and that any solution must be out-of-core, that is, must use disks as storage for temporary data.

It is worth noting that the cost of moving data has lead to the development of so called "communication-avoiding algorithms" [60]. Our approach is orthogonal to these efforts: in communication-avoiding algorithms, the application is modified, in particular some redundant work is done, in order to get rid of some communication operations, whereas in our approach, we do not modify the application, which is provided as a task graph, but we minimize the needed memory peak only by carefully scheduling tasks.

3.3. High-performance computing and linear algebra

Our work on high-performance computing and linear algebra is organized along three research directions. The first direction is devoted to direct solvers of sparse linear systems. The second direction is devoted to combinatorial scientific computing, that is, the design of combinatorial algorithms and tools that solve problems encountered in some of the other research themes, like the problems faced in the preprocessing phases of sparse direct solvers. The last direction deals with the adaptation of classical dense linear algebra kernels to the architecture of future computing platforms.

3.3.1. Direct solvers for sparse linear systems

The solution of sparse systems of linear equations (symmetric or unsymmetric, often with an irregular structure, from a few hundred thousand to a few hundred million equations) is at the heart of many scientific applications arising in domains such as geophysics, structural mechanics, chemistry, electromagnetism, numerical optimization, or computational fluid dynamics, to cite a few. The importance and diversity of applications are a main motivation to pursue research on sparse linear solvers. Because of this wide range of applications, any significant progress on solvers will have a significant impact in the world of simulation. Research on sparse direct solvers in general is very active for the following main reasons:

- many applications fields require large-scale simulations that are still too big or too complicated with respect to today's solution methods;
- the current evolution of architectures with massive, hierarchical, multicore parallelism imposes to overhaul all existing solutions, which represents a major challenge for algorithm and software development;
- the evolution of numerical needs and types of simulations increase the importance, frequency, and size of certain classes of matrices, which may benefit from a specialized processing (rather than resort to a generic one).

Our research in the field is strongly related to the software package MUMPS, which is both an experimental platform for academics in the field of sparse linear algebra, and a software package that is widely used in both academia and industry. The software package MUMPS enables us to (i) confront our research to the real world, (ii) develop contacts and collaborations, and (iii) receive continuous feedback from real-life applications, which is extremely critical to validate our research work. The feedback from a large user community also enables us to direct our long-term objectives towards meaningful directions.

In this context, we aim at designing parallel sparse direct methods that will scale to large modern platforms, and that are able to answer new challenges arising from applications, both efficiently—from a resource consumption point of view—and accurately—from a numerical point of view. For that, and even with increasing parallelism, we do not want to sacrifice in any manner numerical stability, based on threshold partial pivoting, one of the main originalities of our approach (our "trademark") in the context of direct solvers for distributed-memory computers; although this makes the parallelization more complicated, applying the same pivoting strategy as in the serial case ensures numerical robustness of our approach, which we generally measure in terms of sparse backward error. In order to solve the hard problems resulting from the always-increasing demands in simulations, special attention must also necessarily be paid to memory usage (and not only execution time). This requires specific algorithmic choices and scheduling techniques. From a complementary point of view, it is also necessary to be aware of the functionality requirements from the applications and from the users, so that robust solutions can be proposed for a wide range of applications.

Among direct methods, we rely on the multifrontal method [58], [59], [63]. This method usually exhibits a good data locality and hence is efficient in cache-based systems. The task graph associated with the multifrontal method is in the form of a tree whose characteristics should be exploited in a parallel implementation.

Our work is organized along two main research directions. In the first one we aim at efficiently addressing new architectures that include massive, hierarchical parallelism. In the second one, we aim at reducing the running time complexity and the memory requirements of direct solvers, while controlling accuracy.

3.3.2. Combinatorial scientific computing

Combinatorial scientific computing (CSC) is a recently coined term (circa 2002) for interdisciplinary research at the intersection of discrete mathematics, computer science, and scientific computing. In particular, it refers to the development, application, and analysis of combinatorial algorithms to enable scientific computing applications. CSC's deepest roots are in the realm of direct methods for solving sparse linear systems of equations where graph theoretical models have been central to the exploitation of sparsity, since the 1960s. The general approach is to identify performance issues in a scientific computing problem, such as memory use, parallel speed up, and/or the rate of convergence of a method, and to develop combinatorial algorithms and models to tackle those issues.

Our target scientific computing applications are (i) the preprocessing phases of direct methods (in particular MUMPS), iterative methods, and hybrid methods for solving linear systems of equations, and general sparse matrix and tensor computations; and (ii) the mapping of tasks (mostly the sub-tasks of the mentioned solvers) onto modern computing platforms. We focus on the development and the use of graph and hypergraph models, and related tools such as hypergraph partitioning algorithms, to solve problems of load balancing and task mapping. We also focus on bipartite graph matching and vertex ordering methods for reducing the memory overhead and computational requirements of solvers. Although we direct our attention on these models and algorithms through the lens of linear system solvers, our solutions are general enough to be applied to some other resource optimization problems.

3.3.3. Dense linear algebra on post-petascale multicore platforms

The quest for efficient, yet portable, implementations of dense linear algebra kernels (QR, LU, Cholesky) has never stopped, fueled in part by each new technological evolution. First, the LAPACK library [52] relied on BLAS level 3 kernels (Basic Linear Algebra Subroutines) that enable to fully harness the computing power of a single CPU. Then the SCALAPACK library [51] built upon LAPACK to provide a coarse-grain parallel version, where processors operate on large block-column panels. Inter-processor communications

occur through highly tuned MPI send and receive primitives. The advent of multi-core processors has led to a major modification in these algorithms [53], [66], [61]. Each processor runs several threads in parallel to keep all cores within that processor busy. Tiled versions of the algorithms have thus been designed: dividing large block-column panels into several tiles allows for a decrease in the granularity down to a level where many smaller-size tasks are spawned. In the current panel, the diagonal tile is used to eliminate all the lower tiles in the panel. Because the factorization of the whole panel is now broken into the elimination of several tiles, the update operations can also be partitioned at the tile level, which generates many tasks to feed all cores.

The number of cores per processor will keep increasing in the following years. It is projected that high-end processors will include at least a few hundreds of cores. This evolution will require to design new versions of libraries. Indeed, existing libraries rely on a static distribution of the work: before the beginning of the execution of a kernel, the location and time of the execution of all of its component is decided. In theory, static solutions enable to precisely optimize executions, by taking parameters like data locality into account. At run time, these solutions proceed at the pace of the slowest of the cores, and they thus require a perfect load-balancing. With a few hundreds, if not a thousand, cores per processor, some tiny differences between the computing times on the different cores ("jitter") are unavoidable and irremediably condemn purely static solutions. Moreover, the increase in the number of cores per processor once again mandates to increase the number of tasks that can be executed in parallel.

We study solutions that are part-static part-dynamic, because such solutions have been shown to outperform purely dynamic ones [55]. On the one hand, the distribution of work among the different nodes will still be statically defined. On the other hand, the mapping and the scheduling of tasks inside a processor will be dynamically defined. The main difficulty when building such a solution will be to design lightweight dynamic schedulers that are able to guarantee both an excellent load-balancing and a very efficient use of data locality.

SOCRATE Project-Team

3. Research Program

3.1. Flexible Radio Front-End

These are the research axis as they were proposed at the creation of the Socrate Team.

This axis mainly deals with the radio front-end of software radio terminals. In order to ensure a high flexibility in a global wireless network, each node is expected to offer as many degrees of freedom as possible. For instance, the choice of the most appropriate communication resource (frequency channel, spreading code, time slot,...), the interface standard or the type of antenna are possible degrees of freedom. The *multi-** paradigm denotes a highly flexible terminal composed of several antennas providing MIMO features to enhance the radio link quality, which is able to deal with several radio standards to offer interoperability and efficient relaying, and can provide multi-channel capability to optimize spectral reuse. On the other hand, increasing degrees of freedom can also increase the global energy consumption, therefore for energy-limited terminals a different approach has to be defined.

In this research axis, we expect to demonstrate optimization of flexible radio front-end by fine grain simulations, and also by the design of home made prototypes. Of course, studying all the components deeply would not be possible given the size of the team, we are currently not working in new technologies for DAC/ADC and power amplifiers which are currently studied by hardware oriented teams. The purpose of this axis is to build system level simulation taking into account the state of the art of each key component.

3.2. Multi-User Communications

While the first and the third research axes deal with the optimization of the cognitive radio nodes themselves from system and programming point of view, an important complementary objective is to consider the radio nodes in their environments. Indeed, cognitive radio does not target the simple optimization of point to point transmissions, but the optimization of simultaneous concurrent transmissions. The tremendous development of new wireless applications and standards currently observed calls for a better management of the radio spectrum with opportunistic radio access, cooperative transmissions and interference management. This challenge has been identified as one of the most important issue for 5G to guarantee a better exploitation of the spectrum. In addition, mobile internet is going to support a new revolution that is the *tactile internet*, with real time interactions between the virtual and the real worlds, requiring new communication objectives to be met such as low latency end to end communications, distributed learning techniques, in-the-network computation, and many more. The future network will be heterogeneous in terms of technologies, type of data flows and QoS requirements. To address this revolution two work directions have naturally formed within the axis. The first direction concerns the theoretical study of fundamental limits in wireless networks. Introduced by Claude Shannon in the 50s and heavily developed up to today, Information Theory has provided a theoretical foundation to study the performance of wireless communications, not from a practical design view point, but using the statistical properties of wireless channels to establish the fundamental trade-offs in wireless communications. Beyond the classical *energy efficiency - spectral efficiency* tradeoff, information theory and its many derivations, i.e., network information theory, may also help to address additional questions such as determining the optimal rates under decentralized policies, asymptotic behavior when the density of nodes increases, latency controled communication with finite block-length theory, etc. In these cases, information theory is often associated to other theoretical tools such as game theory, stochastic geometry, control theory, graph theory and many others.

Our first research direction consists in evaluating specific mulit-user scenarios from a network information theory perspective, inspired by practical scenarios from various applicative frameworks (e.g. 5G, Wifi, sensor networks, IoT, etc.), and to establish fundamental limits for these scenarios. The second research direction is related to algorithmic and protocol design (PHY/MAC), applied to practical scenarios. Exploiting signal processing, linear algebra inspired models and distributed algorithms, we develop and evaluate various distributed algorithms allowing to improve many QoS metrics such as communication rates, reliability, stability, energy efficiency or computational complexity.

It is clear that both research directions are symbiotic with respect to each other, with the former providing theoretical bounds that serves as a reference to the performance of the algorithms created in the later. In the other way around, the later offers target scenarios for the former, through identifying fundamental problems that are interesting to be studied from the fundamental side. Our contributions of the year in these two directions are summarized further in the document.

3.3. Software Radio Programming Model

Finally the third research axis is concerned with software aspect of the software radio terminal. We have currently two actions in this axis, the first one concerns the programming issues in software defined radio devices, the second one focusses on low power devices: how can they be adapted to integrate some reconfigurability.

The expected contributions of Socrate in this research axis are :

- The design and implementation of a “middleware for SDR”, probably based on a Virtual Machine.
- Prototype implementations of novel software radio systems, using chips from Leti and/or Lyrtech software radio boards.
- Development of a *smart node*: a low-power Software-Defined Radio node adapted to WSN applications.
- Methodology clues and programming tools to program all these prototypes.

3.4. Evolution of the Socrate team

In 2018 the Socrate team which was originaly conceived to develop software defined radio has decided to split in two teams: the Maracas team will consist of the activities of Socrate Axis 2 and be directed by Jean-Marie Gorce, and the Socrate team which will consist in the Axis 1 and 3 of the current version of Socrate. This change is explicit since september 2018 as the Maracas team is created.

The advent of non-volatile memory technologies (NVRAM) is causing a major evolution in all software layers. On the one hand, the non-volatility of data in the event of a breakdown necessarily leads to fatal inconsistencies if the memory is not managed correctly. On the other hand, these memories have very different performances from the usual DRAM, which tends to the appearance of hybrid and complex memory hierarchies. Many technological and scientific challenges are to be faced in all software layers to deal with these two sets of issues. Above all, the answers to be provided depend on the calculation system considered and for what purpose it is constructed.

Within the framework of very low consumption sensors and devices, the Socrate team proposed, with Sytare [28], a software solution allowing to develop embedded applications on platforms supporting an intermittent power supply (TPS – *Transiently Powered System*) and integrating NVRAM as illustrated in Figure 3 . The IPL ZEP (https://project.inria.fr/iplzep/) was also launched by Socrate last year to respond to various scientific challenges related to this issue.

The recent advance in harvesting technologies provides new research direction to Socrate which have skills in radio propagation and low power radio (wake-up radio for instance [31]. Fig 3 , illustrates the *future ultra-low sensor* as envisioned by Socrate.

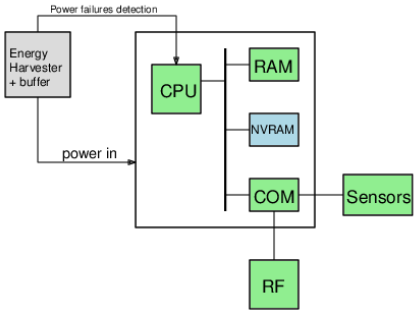


Figure 3. Architecture targeted by Socrate: low energy wireless sensor with peripherals and non volatile memory

SPIRALS Project-Team

3. Research Program

3.1. Introduction

Our research program on self-adaptive software targets two key properties that are detailed in the remainder of this section: *self-healing* and *self-optimization*.

3.2. Objective #1: Self-healing - Mining software artifacts to automatically evolve systems

Software systems are under the pressure of changes all along their lifecycle. Agile development blurs the frontier between design and execution and requires constant adaptation. The size of systems (millions of lines of code) multiplies the number of bugs by the same order of magnitude. More and more systems, such as sensor network devices, live in "surviving" mode, in the sense that they are neither rebootable nor upgradable.

Software bugs are hidden in source code and show up at development-time, testing-time or worse, once deployed in production. Except for very specific application domains where formal proofs are achievable, bugs can not be eradicated. As an order of magnitude, on 16 Dec 2011, the Eclipse bug repository contains 366 922 bug reports. Software engineers and developers work on bug fixing on a daily basis. Not all developers spend the same time on bug fixing. In large companies, this is sometimes a full-time role to manage bugs, often referred to as *Quality Assurance* (QA) software engineers. Also, not all bugs are equal, some bugs are analyzed and fixed within minutes, others may take months to be solved [79].

In terms of research, this means that: (i) one needs means to automatically adapt the design of the software system through automated refactoring and API extraction, (ii) one needs approaches to automate the process of adapting source code in order to fix certain bugs, (iii) one needs to revisit the notion of error-handling so that instead of crashing in presence of errors, software adapts itself to continue with its execution, *e.g.*, in degraded mode.

There is no one-size-fits-all solution for each of these points. However, we think that novel solutions can be found by using **data mining and machine learning techniques tailored for software engineering** [80]. This body of research consists of mining some knowledge about a software system by analyzing the source code, the version control systems, the execution traces, documentation and all kinds of software development and execution artifacts in general. This knowledge is then used within recommendation systems for software development, auditing tools, runtime monitors, frameworks for resilient computing, etc.

The novelty of our approach consists of using and tailoring data mining techniques for analyzing software artifacts (source code, execution traces) in order to achieve the **next level of automated adaptation** (*e.g.*, automated bug fixing). Technically, we plan to mix unsupervised statistical learning techniques (*e.g.* frequent item set mining) and supervised ones (*e.g.* training classifiers such as decision trees). This research is currently not being performed by data mining research teams since it requires a high level of domain expertise in software engineering, while software engineering researchers can use off-the-shelf data mining libraries, such as Weka [58].

We now detail the two directions that we propose to follow to achieve this objective.

3.2.1. Learning from software history how to design software and fix bugs

The first direction is about mining techniques in software repositories (*e.g.*, CVS, SVN, Git). Best practices can be extracted by data mining source code and the version control history of existing software systems. The design and code of expert developers significantly vary from the artifacts of novice developers. We will learn to differentiate those design characteristics by comparing different code bases, and by observing the semantic refactoring actions from version control history. Those design rules can then feed the test-develop-refactor constant adaptation cycle of agile development.

Fault localization of bugs reported in bug repositories. We will build a solid foundation on empirical knowledge about bugs reported in bug repository. We will perform an empirical study on a set of representative bug repositories to identify classes of bugs and patterns of bug data. For this, we will build a tool to browse and annotate bug reports. Browsing will be helped with two kinds of indexing: first, the tool will index all textual artifacts for each bug report; second it will index the semantic information that is not present by default in bug management software—*i.e.*, "contains a stacktrace"). Both indexes will be used to find particular subsets of bug reports, for instance "all bugs mentioning invariants and containing a stacktrace". Note that queries with this kind of complexity and higher are mostly not possible with the state-of-the-art of bug management software. Then, analysts will use annotation features to annotate bug reports. The main outcome of the empirical study will be the identification of classes of bugs that are appropriate for automated localization. Then, we will run machine learning algorithms to identify the latent links between the bug report content and source code features. Those algorithms would use as training data the existing traceability links between bug reports and source code modifications from version control systems. We will start by using decision trees since they produce a model that is explicit and understandable by expert developers. Depending on the results, other machine learning algorithms will be used. The resulting system will be able to locate elements in source code related to a certain bug report with a certain confidence.

Automated bug fix generation with search-based techniques. Once a location in code is identified as being the cause of the bug, we can try to automatically find a potential fix. We envision different techniques: (1) infer fixes from existing contracts and specifications that are violated; (2) infer fixes from the software behavior specified as a test suite; (3) try different fix types one-by-one from a list of identified bug fix patterns; (4) search fixes in a fix space that consists of combinations of atomic bug fixes. Techniques 1 and 2 are explored in [54] and [78]. We will focus on the latter techniques. To identify bug fix patterns and atomic bug fixes, we will perform a large-scale empirical study on software changes (also known as changesets when referring to changes across multiple files). We will develop tools to navigate, query and annotate changesets in a version control system. Then, a grounded theory will be built to master the nature of fixes. Eventually, we will decompose change sets in atomic actions using clustering on changeset actions. We will then use this body of empirical knowledge to feed search-based algorithms (*e.g.* genetic algorithms) that will look for meaningful fixes in a large fix space. To sum up, our research on automated bug fixing will try not only to point to source code locations responsible of a bug, but to search for code patterns and snippets that may constitute the skeleton of a valid patch. Ultimately, a blend of expert heuristics and learned rules will be able to produce valid source code that can be validated by developers and committed to the code base.

3.2.2. Run-time self-healing

The second proposed research direction is about inventing a self-healing capability at run-time. This is complementary to the previous objective that mainly deals with development time issues. We will achieve this in two steps. First, we want to define frameworks for resilient software systems. Those frameworks will help to maintain the execution even in the presence of bugs—*i.e.* to let the system survive. As exposed below, this may mean for example to switch to some degraded modes. Next, we want to go a step further and to define solutions for automated runtime repair, that is, not simply compensating the erroneous behavior, but also determining the correct repair actions and applying them at run-time.

Mining best effort values. A well-known principle of software engineering is the "fail-fast" principle. In a nutshell, it states that as soon as something goes wrong, software should stop the execution before entering incorrect states. This is fine when a human user is in the loop, capable of understanding the error or at least rebooting the system. However, the notion of "failure-oblivious computing" [71] shows that in certain domains, software should run in a resilient mode (*i.e.* capable of recovering from errors) and/or best-effort mode—*i.e.* a slightly imprecise computation is better than stopping. Hence, we plan to investigate data mining techniques in order to learn best-effort values from past executions (*i.e.* somehow learning what is a correct state, or the opposite what is not a completely incorrect state). This knowledge will then be used to adapt the software state and flow in order to mitigate the error consequences, the exact opposite of fail-fast for systems with long-running cycles.

Embedding search based algorithms at runtime. Harman recently described the field of search-based software engineering [59]. We believe that certain search based approaches can be embedded at runtime with the goal of automatically finding solutions that avoid crashing. We will create software infrastructures that allow automatically detecting and repairing faults at run-time. The methodology for achieving this task is based on three points: (1) empirical study of runtime faults; (2) learning approaches to characterize runtime faults; (3) learning algorithms to produce valid changes to the software runtime state. An empirical study will be performed to analyze those bug reports that are associated with runtime information (*e.g.* core dumps or stacktraces). After this empirical study, we will create a system that learns on previous repairs how to produce small changes that solve standard runtime bugs (*e.g.* adding an array bound check to throw a handled domain exception rather than a spurious language exception). To achieve this task, component models will be used to (1) encapsulate the monitoring and reparation meta-programs in appropriate components and (2) support runtime code modification using scripting, reflective or bytecode generation techniques.

3.3. Objective #2: Self-optimization - Sharing runtime behaviors to continuously adapt software

Complex distributed systems have to seamlessly adapt to a wide variety of deployment targets. This is due to the fact that developers cannot anticipate all the runtime conditions under which these systems are immersed. A major challenge for these software systems is to develop their capability to continuously reason about themselves and to take appropriate decisions and actions on the optimizations they can apply to improve themselves. This challenge encompasses research contributions in different areas, from environmental monitoring to real-time symptoms diagnosis, to automated decision making. The variety of distributed systems, the number of optimization parameters, and the complexity of decisions often resign the practitioners to design monolithic and static middleware solutions. However, it is now globally acknowledged that the development of dedicated building blocks does not contribute to the adoption of sustainable solutions. This is confirmed by the scale of actual distributed systems, which can—for example—connect several thousands of devices to a set of services hosted in the Cloud. In such a context, the lack of support for smart behaviors at different levels of the systems can inevitably lead to its instability or its unavailability. In June 2012, an outage of Amazon's Elastic Compute Cloud in North Virginia has taken down Netflix, Pinterest, and Instagram services. During hours, all these services failed to satisfy their millions of customers due to the lack of integration of a self-optimization mechanism going beyond the boundaries of Amazon.

The research contributions we envision within this area will therefore be organized as a reference model for engineering **self-optimized distributed systems** autonomously driven by *adaptive feedback control loops*, which will automatically enlarge their scope to cope with the complexity of the decisions to be taken. This solution introduces a multi-scale approach, which first privileges local and fast decisions to ensure the homeostasis [0] property of a single node, and then progressively propagates symptoms in the network in order to reason on a longer term and a larger number of nodes. Ultimately, domain experts and software developers can be automatically involved in the decision process if the system fails to find a satisfying solution. The research program for this objective will therefore focus on the study of mechanisms for **monitoring, taking decisions, and automatically reconfiguring software at runtime and at various scales**. As stated in the self-healing objective, we believe that there is no one-size-fits-all mechanism that can span all the scales of the system. We will therefore study and identify an optimal composition of various adaptation mechanisms in order to produce long-living software systems.

The novelty of this objective is to exploit the wisdom of crowds to define new middleware solutions that are able to continuously adapt software deployed in the wild. We intend to demonstrate the applicability of this approach to distributed systems that are deployed from mobile phones to cloud infrastructures. The key scientific challenges to address can be summarized as follows: *How does software behave once deployed in the wild? Is it possible to automatically infer the quality of experience, as it is perceived by users? Can the*

[0] Homeostasis is the property of a system that regulates its internal environment and tends to maintain a stable, relatively constant condition of properties [Wikipedia].

runtime optimizations be shared across a wide variety of software? How optimizations can be safely operated on large populations of software instances?

The remainder of this section further elaborates on the opportunities that can be considered within the frame of this objective.

3.3.1. Monitoring software in the wild

Once deployed, developers are generally no longer aware of how their software behave. Even if they heavily use testbeds and benchmarks during the development phase, they mostly rely on the bugs explicitly reported by users to monitor the efficiency of their applications. However, it has been shown that contextual artifacts collected at runtime can help to understand performance leaks and optimize the resilience of software systems [81]. Monitoring and understanding the context of software at runtime therefore represent the first building block of this research challenge. Practically, we intend to investigate crowd-sensing approaches, to smartly collect and process runtime metrics (*e.g.*, request throughput, energy consumption, user context). Crowd-sensing can be seen as a specific kind of crowdsourcing activity, which refers to the capability of lifting a (large) diffuse group of participants to delegate the task of retrieving trustable data from the field. In particular, crowd-sensing covers not only *participatory sensing* to involve the user in the sensing task (*e.g.*, surveys), but also *opportunistic sensing* to exploit mobile sensors carried by the user (*e.g.*, smartphones).

While reported metrics generally enclose raw data, the monitoring layer intends to produce meaningful indicators like the *Quality of Experience* (QoE) perceived by users. This QoE reflects representative symptoms of software requiring to trigger appropriate decisions in order to improve its efficiency. To diagnose these symptoms, the system has to process a huge variety of data including runtime metrics, but also history of logs to explore the sources of the reported problems and identify opportunities for optimizations. The techniques we envision at this level encompass machine learning, principal component analysis, and fuzzy logic [70] to provide enriched information to the decision level.

3.3.2. Collaborative decision-making approaches

Beyond the symptoms analysis, decisions should be taken in order to improve the *Quality of Service* (QoS). In our opinion, collaborative approaches represent a promising solution to effectively converge towards the most appropriate optimization to apply for a given symptom. In particular, we believe that exploiting the wisdom of the crowd can help the software to optimize itself by sharing its experience with other software instances exhibiting similar symptoms. The intuition here is that the body of knowledge that supports the optimization process cannot be specific to a single software instance as this would restrain the opportunities for improving the quality and the performance of applications. Rather, we think that any software instance can learn from the experience of others.

With regard to the state-of-the-art, we believe that a multi-levels decision infrastructure, inspired from distributed systems like Spotify [57], can be used to build a decentralized decision-making algorithm involving the surrounding peers before requesting a decision to be taken by more central control entity. In the context of collaborative decision-making, peer-based approaches therefore consist in quickly reaching a consensus on the decision to be adopted by a majority of software instances. Software instances can share their knowledge through a micro-economic model [51], that would weight the recommendations of experienced instances, assuming their age reflects an optimal configuration.

Beyond the peer level, the adoption of algorithms inspired from evolutionary computations, such as genetic programming, at an upper level of decision can offer an opportunity to test and compare several alternative decisions for a given symptom and to observe how does the crowd of applications evolves. By introducing some diversity within this population of applications, some instances will not only provide a satisfying QoS, but will also become naturally resilient to unforeseen situations.

3.3.3. Smart reconfigurations in the large

Any decision taken by the crowd requires to propagate back to and then operated by the software instances. While simplest decisions tend to impact software instances located on a single host (*e.g.*, laptop, smartphone),

this process can also exhibit more complex reconfiguration scenarios that require the orchestration of various actions that have to be safely coordinated across a large number of hosts. While it is generally acknowledged that centralized approaches raise scalability issues, we think that self-optimization should investigate different reconfiguration strategies to propagate and apply the appropriate actions. The investigation of such strategies can be addressed in two steps: the consideration of *scalable data propagation protocols* and the identification of *smart reconfiguration mechanisms*.

With regard to the challenge of scalable data propagation protocols, we think that research opportunities encompass not only the exploitation of gossip-based protocols [56], but also the adoption of publish/subscribe abstractions [64] in order to decouple the decision process from the reconfiguration. The fundamental issue here is the definition of a communication substrate that can accommodate the propagation of decisions with relaxed properties, inspired by *Delay Tolerant Networks* (DTN), in order to reach weakly connected software instances. We believe that the adoption of asynchronous communication protocols can provide the sustainable foundations for addressing various execution environments including harsh environments, such as developing countries, which suffer from a partial connectivity to the network. Additionally, we are interested in developing the principle of *social networks of applications* in order to seamlessly group and organize software instances according to their similarities and acquaintances. The underlying idea is that grouping application instances can contribute to the identification of optimization profiles not only contributing to the monitoring layer, but also interested in similar reconfigurations. Social networks of applications can contribute to the anticipation of reconfigurations by exploiting the symptoms of similar applications to improve the performance of others before that problems actually happen.

With regard to the challenge of smart reconfiguration mechanisms, we are interested in building on our established experience of adaptive middleware [75] in order to investigate novel approaches to efficient application reconfigurations. In particular, we are interested in adopting seamless micro-updates and micro-reboot techniques to provide in-situ reconfiguration of pieces of software. Additionally, the provision of safe and secured reconfiguration mechanisms is clearly a key issue that requires to be carefully addressed in order to avoid malicious exploitation of dynamic reconfiguration mechanisms against the software itself. In this area, although some reconfiguration mechanisms integrate transaction models [65], most of them are restricted to local reconfigurations, without providing any support for executing distributed reconfiguration transactions. Additionally, none of the approached published in the literature include security mechanisms to preserve from unauthorized or malicious reconfigurations.

STACK Project-Team

3. Research Program

3.1. Overview

STACK research activities have been organized around four research topics. The two first ones are related to the resource management mechanisms and the programming support that are mandatory to operate and use ICT geo-distributed resources (compute, storage, network). They are transverse to the System/Middleware/Application layers, which generally composed a software stack, and nurture each other (*i.e.*, the resource management mechanisms will leverage abstractions/concepts proposed by the programming support axis and reciprocally). The third and fourth research topics are related to the Energy and Security dimensions (both also crosscutting the three software layers). Although they could have been merged with the two first axes, we identified them as independent research directions due to their critical aspects with respect to the societal challenges they represent. In the following, we detail the actions we plan to do in each research direction.

3.2. Resource Management

The challenge in this axis is to identify, design or revise mechanisms that are mandatory to operate and use a set of massively geo-distributed resources in an efficient manner [50]. This covers considering challenges at the scale of nodes, within one site (*i.e.*, one geographical location) and throughout the whole geo-distributed ICT infrastructure. It is noteworthy that the network community has been investigating similar challenges for the last few years [69]. To benefit from their expertise, in particular on how to deal with intermittent networks, STACK members have recently initiated exchanges and collaborative actions with some network research groups and telcos (see Sections 8.1 and 9.1). We emphasize, however, that we do not deliver contributions related to network equipments/protocols. The scientific and technical achievements we aim to deliver are related to the (distributed) system aspects.

3.2.1. Performance Characterization of Low-Level Building Blocks

Although Cloud Computing has enabled the consolidation of services and applications into a subset of servers, current operating system mechanisms do not provide appropriate abstractions to prevent (or at least control) the performance degradation that occurs when several workloads compete for the same resources [101]. Keeping in mind that server density is going to increase with physical machines composed of more and more cores and that applications will be more and more data intensive, it is mandatory to identify interferences that appear at a low level on each dimension (compute, memory, network, and storage) and propose countermeasures. In particular, previous studies [101], [61] on pros and cons of current technologies – virtual machines (VMs) [75], [83], containers and microservices – which are used to consolidate applications on the same server, should be extended: In addition to evaluating the performance we can expect from each of these technologies on a single node, it is important to investigate interferences that may result from cross-layer and remote communications [102]. We will consider in particular all interactions related to geo-distributed systems mechanisms/services that are mandatory to operate and use geo-distributed ICT infrastructures.

3.2.2. Geo-Distributed System Mechanisms

Although several studies have been highlighting the advantages of geo-distributed ICT infrastructures in various domains (see Section 3.), progress on how to operate and use such infrastructures is marginal. Current solutions [32] [33] are rather close to the initial Cisco Fog Computing proposal that only allows running domain-specific applications on edge resources and centralized Cloud platforms [40] (in other words, these solutions do not allow running stateful workloads in isolated environments such as containers or VMs). More recently, solutions leveraging the idea of federating VIMs (as the aforementioned ETSI MEC proposal [88]) have been proposed. ONAP [95], an industry-driven solution, enables the orchestration and

automation of virtual network functions across distinct VIMs. From the academic side, FogBow [42] aims to support federations of Infrastructure-as-a-Service (IaaS) providers. Finally, NIST initiated a collaborative effort with IEEE to advance Federated Cloud platforms through the development of a conceptual architecture and a vocabulary[0]. Although all these projects provide valuable contributions, they face the aforementioned orchestration limitations (*i.e.*, they do not manage decisions taken in each VIM). Moreover, they all have been designed by only considering the developer/user's perspective. They provide abstractions to manage the life cycle of geo-distributed applications, but do not address administrative requirements.

To cope with specifics of Wide-Area networks while delivering most features that made Cloud Computing solutions successful also at the edge, our community should first identify limitations/drawbacks of current resource management system mechanisms with respect to the Fog/Edge requirements and propose revisions when needed [68], [81].

To achieve this aim, STACK members propose to conduct first a series of studies aiming at understanding the software architecture and footprint of major services that are mandatory for operating and using Fog/Edge infrastructures (storage backends, monitoring services, deployment/reconfiguration mechanisms, etc.). Leveraging these studies, we will investigate how these services should be deployed in order to deal with resources constraints, performance variability, and network split brains. We will rely on contributions that have been accomplished in distributed algorithms and self-* approach for the last decade. In the short and medium term, we plan to evaluate the relevance of NewSQL systems [53] to store internal states of distributed system mechanisms in an edge context, and extend our proposals on new storage backends such as key/value stores [52], [94], and burst buffers [103]. We also plan to conduct new investigations on data-stream frameworks for Fog and Edge infrastructures [47]. These initial contributions should enable us to identify general rules to deliver other advanced system mechanisms that will be mandatory at the higher levels in particular for the deployment and reconfiguration manager in charge of orchestrating all resources.

3.2.3. Capacity Planning and Placement Strategies

An objective shared by users and providers of ICT infrastructures is to limit as much as possible the operational costs while providing the expected and requested quality of service (QoS). To optimize this cost while meeting QoS requirements, data and applications have to be placed in the best possible way onto physical resources according to data sources, data types (stream, graphs), application constraints (real-time requirements) and objective functions. Furthermore, the placement of applications must evolve through time to cope with the fluctuations in terms of application resource needs as well as the physical events that occur at the infrastructure level (resource creation/removals, hardware failures, etc.). This placement problem, *a.k.a.* the deployment and reconfiguration challenge as it will be described in Section 3.3 , can be modelized in many different ways, most of the time by multi-dimensional and multi-objective bin-packing problems or by scheduling problems which are known to be difficult to solve. Many studies have been performed, for example, to optimize the placement of virtual machines onto ICT infrastructures [77]. STACK will inherit the knowledge acquired through previous activities in this domain, particularly its use of constraint programming strategies in autonomic managers [73], [72], relying on MAPE (monitor, analyze, plan, and execute) control loops. While constraint programming approaches are known to hardly scale, they enable the composition of various constraints without requiring to change heuristic algorithms each time a new constraint has to be considered [71]. We believe it is a strong advantage to deal with the diversity brought by geo-distributed ICT infrastructures. Moreover, we have shown in previous work that decentralized approaches can tackle the scalability issue while delivering placement decisions good enough and sometimes close to the optimal [87].

Leveraging this expertise, we propose, first, to identify new constraints raised by massively geo-distributed infrastructures (*e.g.*, data locality, energy, security, reliability and the heterogeneity and mobility of the underlying infrastructure). Based on this preliminary study, we will explore new placement strategies not only for computation sandboxes but for data (location, replication, streams, etc.) in order to benefit from the geo-distribution of resources and meet the required QoS.These investigations should lead to collaborations with operational research and optimization groups such as TASC, another research group from IMT Atlantique.

[0] https://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/FederatedCloudPWGFC (Dec 2018).

Second, we will leverage contributions made on the previous axis "Performance Characterization of Low-Level Building Blocks" to determine how the deployment of the different units (software components and data sets) should be executed in order to reduce as much as possible the time to reconfigure the system (*i.e.*, the *Execution* phase in the control loop). In some recent work [83], we have shown that the provisioning of a new virtual machine should be done carefully to mitigate boot penalties. More generally, proposing an efficient action plan for the *Execution* phase will be a major point as Wide-Area-Network specifics may lead to significant delays, in particular when the amount of data to be manipulated is important.

Finally, we will investigate new approaches to decentralize the placement process while considering the geo-distributed context. Among the different challenges to address, we will study how a combination of autonomic managers, at both the infrastructure and application levels [60], could be proposed in a decentralized manner. Our first idea is to geo-distribute a fleet of small control loops over the whole infrastructure. By improving the locality of data collection and weakening combinatorics, these loops would allow the system to address responsiveness and quality expectations.

3.3. Programming Support

We pursue two main research directions relative to new programming support: first, developing new programming models with appropriate support in existing languages (libraries, embedded DSLs, etc.) and, second, providing new means for deployment and reconfiguration in geo-distributed ICT environments, principally supporting the mapping of software onto the infrastructure. For both directions two levels of challenges are considered. On the one hand, the *generic* level refers to efforts on programming support that can be applied to any kind of distributed software, application or system. On this level, contributions could thus be applied to any of the three layers addressed by STACK (*i.e.*, system, middleware or application). On the other hand, the corresponding generic programming means may not be appropriate in practice (*e.g.*, requirements for more dedicated support, performance constraints, etc.), even if they may lead to interesting general properties. For this reason, a *specific* level is also considered. This level could be based on the generic one but addresses specific cases or domains.

3.3.1. Programming Models and Languages Extensions

The current landscape of programming support for cloud applications is fragmented. This fragmentation is based on apparently different needs for various kinds of applications, in particular, web-based, computation-based, focusing on the organization of the computation, and data-based applications, within the last case a quite strong dichotomy between applications considering data as sets or relations, close to traditional database applications and applications considering data as real-time streams. This has led to various programming models, in a loose sense, including for instance microservices, graph processing, dataflows, streams, etc. These programming models have mostly been offered to the application programmer in the guise of frameworks, each offering subtle variants of the programming models with various implementation decisions favoring particular application and infrastructure settings. Whereas most frameworks are dedicated to a given programming model, *e.g.*, basic Pregel [82], Hive [97], Hadoop [98], some of them are more general-purpose through the provision of several programming models, *e.g.*, Flink [46] and Spark [79]. Finally, some dedicated language support has been considered for some models (*e.g.*, the language SPL underlying IBM Streams [74]) as well as core languages and calculi (*e.g.*, [43], [92]).

This situation raises a number of challenges on its own, related to a better structuring of the landscape. It is necessary to better understand the various programming models and their possible relations, with the aim of facilitating, if not their complete integration, at least their composition, at the conceptual level but also with respect to their implementations, as specific languages and frameworks.

Switching to massively geo-distributed infrastructures adds to these challenges by leading to a new range of applications (*e.g.*, smart-* applications) that, by nature, require mixing these various programming models, together with a much more dynamic management of their runtime.

In this context, STACK would like to explore two directions:

- First, we propose to contribute to generic programming models and languages to address composability of different programming models [55]. For example, providing a generic stream data processing model that can operate under both data stream [46] and operation stream [104] modes, thus streams can be processed in micro batches to favour high throughput or record by record to sustain low latency. Software engineering properties such as separation of concerns and composition should help address such challenges [35], [93]. They should also facilitate the software deployment and reconfiguration challenges discussed below.
- Second, we plan to revise relevant programming models, the associated specific languages, and their implementation according to the massive geo-distribution of the underlying infrastructure, the data sources, and application end-users. For example, although SPL is extensible and distributed, it has been designed to run on multi-cores and clusters [74]. It does not provide the level of dynamicity required by geo-distributed applications (*e.g.*, to handle topology changes, loss of connectivity at the edge, etc.). Moreover, as more network data transfers will happen within a massively geo-distributed infrastructure, correctness of data transfers should be guaranteed. This has potential impact from the programming models to their implementations.

3.3.2. Deployment and Reconfiguration Challenges

The second research direction deals with the complexity of deploying distributed software (whatever the layer, application, middleware or system) onto an underlying infrastructure. As both the deployed pieces of software and the infrastructures addressed by STACK are large, massively distributed, heterogeneous and highly dynamic, the deployment process cannot be handled manually by developers or administrators. Furthermore, and as already mentioned in Section 3.2 , the initial deployment of some distributed software will evolve through time because of the dynamicity of both the deployed software and the underlying infrastructures. When considering reconfiguration, which encompasses deployment as a specific case, the problem becomes more difficult for two main reasons: (1) the current state of both the deployed software and the infrastructure has to be taken into account when deciding on a reconfiguration plan, (2) as the software is already running the reconfiguration should minimize disruption time, while avoiding inconsistencies [80], [85]. Many deployment tools have been proposed both in academia and industry [57]. For example, Ansible [0], Chef [0] and Puppet [0] are very well-known generic tools to automate the deployment process through a set of batch instructions organized in groups (*e.g.*, *playbooks* in Ansible). Some tools are specific to a given environment, like Kolla to deploy OpenStack, or the embedded deployment manager within Spark. Few reconfiguration capabilities are available in production tools such as *scaling* and *restart* after a fault [0] [0]. Academia has contributed to generic deployment and reconfiguration models. Most of these contributions are component-based. Component models divide a distributed software as a set of component instances (or modules) and their assembly, where components are connected through well defined interfaces [93]. Thus, modeling the reconfiguration process consists in describing the life cycle of different components and their interactions. Most component-based approaches offer a fixed life cycle, *i.e.*, identical for any component [62]. Two main contributions are able to customize life cycles, Fractal [45], [38] and its evolutions [35], [36], [59], and Aeolus [54]. In Fractal, the *control* part of a component (*e.g.*, its life cycle) is modeled itself as a component assembly that is highly flexible. Aeolus, on the other hand, offers a finer control on both the evolution and the synchronization of the deployment process by modeling each component life cycle with a finite state machine.

A reconfiguration raises at least five questions, all of them are correlated: (1) *why software has to be reconfigured?* (monitoring, modeling and analysis) (2) *what should be reconfigured?* (software modeling and analysis), (3) *how should it be reconfigured?* (software modeling and planning decisions), (4) *where should it*

[0] https://www.ansible.com/
[0] https://www.chef.io/chef/
[0] https://puppet.com/
[0] https://kubernetes.io/
[0] https://jujucharms.com/

be reconfigured? (infrastructure modeling and planning decisions), and (5) *when to reconfigure it?* (scheduling algorithms). STACK will contribute to all aspects of a reconfiguration process as described above. However, according to the expertise of STACK members, we will focus mainly on the three first questions: *why*, *what* and *how*, leaving questions *where* and *when* to collaborations with operational research and optimization teams.

First of all, we would like to investigate *why software has to be reconfigured?* Many reasons could be mentioned, such as hardware or software fault tolerance, mobile users, dynamicity of software services, etc. All those reasons are related somehow to the Quality of Service (QoS) or the Service Level Agreement (SLA) between the user and the Cloud provider. We first would like to explore the specificities of QoS and SLAs in the case of massively geo-distributed ICT environments [89]. By being able to formalize this question, analyzing the requirement of a reconfiguration will be facilitated.

Second, we think that four important properties should be enhanced when deploying and reconfiguring models in massively geo-distributed ICT environments. First, as low-latency applications and systems will be subject to deployment and reconfiguration, the performance and the ability to scale are important. Second, as many different kinds of deployments and reconfigurations will concurrently hold within the infrastructure, processes have to be reliable, which is facilitated by a fine-grained control of the process. Finally, as many different software elements will be subject to deployment and reconfiguration, common generic models and engines for deployment and reconfiguration should be designed [44]. For these reasons, we intend to go beyond Aeolus by: first, leveraging the expression of parallelism within the deployment process, which should lead to better performance; second, improving the separation of concerns between the component developer and the reconfiguration developer; third, enhancing the possibility to perform concurrent and decentralized reconfigurations.

Research challenges relative to programming support have been presented above. Many of these challenges are related, in different manners, to the resource management level of STACK or to crosscutting challenges, *i.e.*, energy and security. First, one can notice that any programming model or deployment and reconfiguration implementation should be based on mechanisms related to resource management challenges. For this reason, all challenges addressed within this section are linked with lower level building blocks presented in Section 3.2 . Second, as detailed above, deployment and reconfiguration address at least five questions. The question *what?* is naturally related to programming support. However, questions *why*, *how?*, *where?* and *when?* are also related to Section 3.2 , for example, to monitoring and capacity planning. Moreover, regarding the deployment and reconfiguration challenges, one can note that the same goals recursively happen when deploying the control building blocks themselves (bootstrap issue). This comforts the need to design generic deployment and reconfiguration models and frameworks. These low-level models should then be used as back-ends to higher-level solutions. Finally, as *energy* and *security* are crosscutting themes within the STACK project, many additional energy and security considerations could be added to the above challenges. For example, our deployment and reconfiguration frameworks and solutions could be used to guarantee the deployment of end-to-end security policies or to answer specific energy constraints [70] as detailed in the next section.

3.4. Energy

The overall electrical consumption of DCs grows according to the demand of Utility Computing. Considering that the latter has been continuously increasing since 2008, the energy footprint of Cloud services overall is nowadays critical with about 91 billion kilowatt-hours of electricity [91]. Besides the ecological impact, the energy consumption is a predominant criterion for providers since it determines a large part of the operational cost of their infrastructure. Among the different appraoches that have been investigated to reduce the energy footprint, some studies have been ingestigating the use of renewable energy sources to power microDCs [64]. Workload distribution for geo-distributed DCs is also another promising approach [66], [78], [99]. Our research will extend these results with the ultimate goal of considering the different opportunities to control the energy footprint across the whole stack (hardware and software opportunities, renewable energy, thermal management, etc.). In particular, we identified several challenges that we will address in this context within the STACK framework.

First, we propose to evaluate the energy efficiency of low-level building blocks, from the viewpoints of computation (VMs, containers, microkernel, microservices) [58] and data (hard drives, SSD, in-memory storage, distributed file systems). For computations, in the continuity of our previous work [56], [73], we will investigate workload placement policies according to energy (minimizing energy consumption, power capping, thermal load balancing, etc.). Regarding the data dimension, we will investigate, in particular, the trade-offs between energy consumption and data availability, durability and consistency [51], [94]. Our ambition is to propose an adaptive energy-aware data layout and replication scheme to ensure data availability with minimal energy consumption. It is noteworthy that these new activities will also consider our previous work on DCs partially powered by renewable energy (see the SeDuCe project, in Section 6.7), with the ultimate goal of reducing the CO_2 footprint.

Second, we will complete current studies to understand pros and cons of massively geo-distributed infrastructures from the energy perspective. Addressing the energy challenge is a complex task that involves considering several dimensions such as the energy consumption due to the physical resources (CPU, memory, disk, network), the performance of the applications (from the computation and data viewpoints), and the thermal dissipation caused by air conditioning in each DC. Each of these aspects can be influenced by each level of the software stack (*i.e.*, low-level building blocks, coordination and autonomous loops, and finally application life cycle). In previous projects, we have studied and modeled the consumption of the main components, notably the network, as part of a single microDC. We plan to extend these models to deal with geo-distribution. The objective is to propose models that will enable us to refine our placement algorithms as discussed in the next paragraph. These models should be able to consider the energy consumption induced by all WAN data exchanges, including site-to-site data movements as well as the end users' communications for accessing virtualized resources.

Third, we expect to implement green-energy-aware balancing strategies, leveraging the aforementioned contributions. Although the infrastructures we envision increase complexity (because WAN aspects should also be taken into account), the geo-distribution of resources brings several opportunities from the energy viewpoint. For instance, it is possible to define several workload/data placement policies according to renewable energy availability. Moreover, a tightly-coupled software stack allows users to benefit from such a widely distributed infrastructure in a transparent way while enabling administrators to balance resources in order to benefit from green energy sources when available. An important difficulty, compared to centralized infrastructures, is related to data sharing between software instances. In particular, we will study issues raised by the distribution and replication of services across several microDCs. In this new context, many challenges must be addressed: where to place the data (Cloud, Edge) in order to mitigate dat a movements? What is the impact in terms of energy consumption, network and response time of these two approaches? How to manage the consistency of replicated data/services? All these aspects must be studied and integrated into our placement algorithms.

Fourth, we will investigate the energy footprint of the current techniques that address failure and performance variability in large-scale systems. For instance, *stragglers* (*i.e.*, tasks that take a significantly longer time to finish than the normal execution time) are natural results of performance variability, they cause extra resource and energy consumption. Our goal is to understand the energy overhead of these techniques and introduce new handling techniques that take into consideration the energy efficiency of the platform [86].

Finally, in order to answer specific energy constraints, we want to reify energy aspects at the application level and propose a metric related to the use of energy (Green SLA [34]), for example to describe the maximum allowed CO_2 emissions of a Fog/Edge service. Unlike other approaches [67], [39], [65] that attempt to identify the best trade-off, we want to offer to developers/end-users the opportunity to select the best choice between application performance, correctness and energy footprint. Such a capability will require reifying the energy dimension at the level of big-data and interactive applications. Besides, with the emergence of renewable energy (*e.g.*, solar panels for microDC), investigating the energy consumption *vs* performance trade-off [70] and the smart usage of green energy for ICT geo-distributed services seems promising. For example, we want to offer the opportunity to developers/end-users to control the scaling of the applications based on this trade-

off instead of current approaches that only considered application load. Providing such a capability will also require appropriate software abstractions.

3.5. Security

Because of its large size and complex software structure, geo-distributed applications and infrastructures are particularly exposed to security and privacy issues [90]. They are subject to numerous security vulnerabilities that are frequently exploited by malicious attackers in order to exfiltrate personal, institutional or corporate data. Securing these systems require security and privacy models and corresponding techniques that are applicable at all software layers in order to guard interactions at each level but also between levels. However, very few security models exist for the lower layers of the software stack and no model enables the handling of interactions involving the complete software stack. Any modification to its implementation, deployment status, configuration, etc., may introduce new or trigger existing security and privacy issues. Finally, applications that execute on top of the software stack may introduce security issues or be affected by vulnerabilities of the stack. Overall, security and privacy issues are therefore interdependent with all other activities of the STACK team and constitute an important research topic for the team.

As part of the STACK activities, we consider principally security and privacy issues related to the vertical and horizontal compositions of software components forming the software stack and the distributed applications running on top of it. Modifications to the *vertical composition* of the software stack affect different software levels at once. As an example, side-channel attacks often target virtualized services (*i.e.*, services running within VMs); attackers may exploit insecure hardware caches at the system level to exfiltrate data from computations at the higher level of VM services [84], [100]. Security and privacy issues also affect *horizontal compositions*, that is, compositions of software abstractions on one level: most frequently horizontal compositions are considered on the level of applications/services but they are also relevant on the system level or the middleware level, such as compositions involving encryption and database fragmentation services.

The STACK members aim at addressing two main research issues: enabling full-stack (vertical) security and per-layer (horizontal) security. Both of these challenges are particularly hard in the context of large geo-distributed systems because they are often executed on heterogeneous infrastructures and are part of different administrative domains and governed by heterogeneous security and privacy policies. For these reasons they typically lack centralized control, are frequently subject to high latency and are prone to failures.

Concretely, we will consider two classes of security and privacy issues in this context. First, on a general level, we strive for a method for the programming and reasoning about compositions of security and privacy mechanisms including, but not limited to, encryption, database fragmentation and watermarking techniques. Currently, no such general method exists, compositions have only been devised for specific and limited cases, for example, compositions that support the commutation of specific encryption and watermarking techniques [76], [48]. We provided preliminary results on such compositions [49] and have extended them to biomedical, notably genetic, analyses in the e-health domain [41]. Second, on the level of security and privacy properties, we will focus on isolation properties that can be guaranteed through vertical and horizontal composition techniques. We have proposed first results in this context in form of a compositional notion of distributed side channel attacks that operate on the system and middleware levels [37].

It is noteworthy that the STACK members do not have to be experts on the individual security and privacy mechanisms, such as watermarking and database fragmentation. We are, however, well-versed in their main properties so that we can integrate them into our composition model. We also interact closely with experts in these techniques and the corresponding application domains, notably e-health for instance, in the context of the PRIVGEN project[0], see Section 9.1 .

More generally, we highlight that security issues in distributed systems are very closely related to the other STACK challenges, dimensions and research directions. Guaranteeing security properties across the software stack and throughout software layers in highly volatile and heterogeneous geo-distributed systems is expected to harness and contribute results to the self-management capabilities investigated as part of the team's resource

[0]Privacy-preserving sharing and processing of genetic data, https://privgen.cominlabs.u-bretagneloire.fr/fr

management challenges. Furthermore, security and privacy properties are crosscutting concerns that are intimately related to the challenges of application life cycle management. Similarly, the security issues are also closely related to the team's work on programming support. This includes new means for programming, notably in terms of event and stream programming, but also the deployment and reconfiguration challenges, notably concerning automated deployment. As a crosscutting functionality, the security challenges introduced above must be met in an integrated fashion when designing, constructing, executing and adapting distributed applications as well as managing distributed resources.

STORM Project-Team

3. Research Program

3.1. Parallel Computing and Architectures

Following the current trends of the evolution of HPC systems architectures, it is expected that future Exascale systems (i.e. Sustaining 10^{18} flops) will have millions of cores. Although the exact architectural details and trade-offs of such systems are still unclear, it is anticipated that an overall concurrency level of $O(10^9)$ threads/tasks will probably be required to feed all computing units while hiding memory latencies. It will obviously be a challenge for many applications to scale to that level, making the underlying system sound like "embarrassingly parallel hardware."

From the programming point of view, it becomes a matter of being able to expose extreme parallelism within applications to feed the underlying computing units. However, this increase in the number of cores also comes with architectural constraints that actual hardware evolution prefigures: computing units will feature extra-wide SIMD and SIMT units that will require aggressive code vectorization or "SIMDization", systems will become hybrid by mixing traditional CPUs and accelerators units, possibly on the same chip as the AMD APU solution, the amount of memory per computing unit is constantly decreasing, new levels of memory will appear, with explicit or implicit consistency management, etc. As a result, upcoming extreme-scale system will not only require unprecedented amount of parallelism to be efficiently exploited, but they will also require that applications generate adaptive parallelism capable to map tasks over heterogeneous computing units.

The current situation is already alarming, since European HPC end-users are forced to invest in a difficult and time-consuming process of tuning and optimizing their applications to reach most of current supercomputers' performance. It will go even worse with the emergence of new parallel architectures (tightly integrated accelerators and cores, high vectorization capabilities, etc.) featuring unprecedented degree of parallelism that only too few experts will be able to exploit efficiently. As highlighted by the ETP4HPC initiative, existing programming models and tools won't be able to cope with such a level of heterogeneity, complexity and number of computing units, which may prevent many new application opportunities and new science advances to emerge.

The same conclusion arises from a non-HPC perspective, for single node embedded parallel architectures, combining heterogeneous multicores, such as the ARM big.LITTLE processor and accelerators such as GPUs or DSPs. The need and difficulty to write programs able to run on various parallel heterogeneous architectures has led to initiatives such as HSA, focusing on making it easier to program heterogeneous computing devices. The growing complexity of hardware is a limiting factor to the emergence of new usages relying on new technology.

3.2. Scientific and Societal Stakes

In the HPC context, simulation is already considered as a third pillar of science with experiments and theory. Additional computing power means more scientific results, and the possibility to open new fields of simulation requiring more performance, such as multi-scale, multi-physics simulations. Many scientific domains able to take advantage of Exascale computers, these "Grand Challenges" cover large panels of science, from seismic, climate, molecular dynamics, theoretical and astrophysics physics... Besides, embedded applications are also able to take advantage of these performance increase. There is still an on-going trend where dedicated hardware is progressively replaced by off-the-shelf components, adding more adaptability and lowering the cost of devices. For instance, Error Correcting Codes in cell phones are still hardware chips, but with the forthcoming 5G protocol, new software and adaptative solutions relying on low power multicores are also explored. New usages are also appearing, relying on the fact that large computing capacities are becoming more affordable and widespread. This is the case for instance with Deep Neural Networks where the training phase can be done

on supercomputers and then used in embedded mobile systems. The same consideration applies for big data problems, of internet of things, where small sensors provide large amount of data that need to be processed in short amount of time. Even though the computing capacities required for such applications are in general a different scale from HPC infrastructures, there is still a need in the future for high performance computing applications.

However, the outcome of new scientific results and the development of new usages for mobile, embedded systems will be hindered by the complexity and high level of expertise required to tap the performance offered by future parallel heterogeneous architectures.

3.3. Towards More Abstraction

As emphasized by initiatives such as the European Exascale Software Initiative (EESI), the European Technology Platform for High Performance Computing (ETP4HPC), or the International Exascale Software Initiative (IESP), the HPC community needs new programming APIs and languages for expressing heterogeneous massive parallelism in a way that provides an abstraction of the system architecture and promotes high performance and efficiency. The same conclusion holds for mobile, embedded applications that require performance on heterogeneous systems.

This crucial challenge given by the evolution of parallel architectures therefore comes from this need to make high performance accessible to the largest number of developers, abstracting away architectural details providing some kind of performance portability, and provided a high level feed-back allowing the user to correct and tune the code. Disruptive uses of the new technology and groundbreaking new scientific results will not come from code optimization or task scheduling, but they require the design of new algorithms that require the technology to be tamed in order to reach unprecedented levels of performance.

Runtime systems and numerical libraries are part of the answer, since they may be seen as building blocks optimized by experts and used as-is by application developers. The first purpose of runtime systems is indeed to provide *abstraction*. Runtime systems offer a uniform programming interface for a specific subset of hardware (e.g., OpenGL or DirectX are well-established examples of runtime systems dedicated to hardware-accelerated graphics) or low-level software entities (e.g., POSIX-thread implementations). They are designed as thin user-level software layers that complement the basic, general purpose functions provided by the operating system calls. Applications then target these uniform programming interfaces in a portable manner. Low-level, hardware dependent details are hidden inside runtime systems. The adaptation of runtime systems is commonly handled through drivers. The abstraction provided by runtime systems thus enables portability. Abstraction alone is however not enough to provide portability of performance, as it does nothing to leverage low-level-specific features to get increased performance and does nothing to help the user tune his code. Consequently, the second role of runtime systems is to *optimize* abstract application requests by dynamically mapping them onto low-level requests and resources as efficiently as possible. This mapping process makes use of scheduling algorithms and heuristics to decide the best actions to take for a given metric and the application state at a given point in its execution time. This allows applications to readily benefit from available underlying low-level capabilities to their full extent without breaking their portability. Thus, optimization together with abstraction allows runtime systems to offer portability of performance. Numerical libraries provide sets of highly optimized kernels for a given field (dense or sparse linear algebra, FFT, etc.) either in an autonomous fashion or using an underlying runtime system.

Application domains cannot resort to libraries for all codes however, computation patterns such as stencils are a representative example of such difficulty. The compiler technology plays here a central role, in managing high level semantics, either through templates, domain specific languages or annotations. Compiler optimizations, and the same applies for runtime optimizations, are limited by the level of semantics they manage. Providing part of the algorithmic knowledge of an application, for instance knowing that it computes a 5-point stencil and then performs a dot product, would lead to more opportunities to adapt parallelism, memory structures, and is a way to leverage the evolving hardware. Besides, with the need for automatic optimization comes the need for *feed-back* to the user, corresponding to the need to debug the code and also to understand what the runtime

has performed. Here the compiler plays also a central role in the analysis of the code, and the instrumentation of the program given to the runtime.

Compilers and runtime play a crucial role in the future of high performance applications, by defining the input language for users, and optimizing/transforming it into high performance code. The objective of STORM is to propose better interactions between compiler and runtime and more semantics for both approaches.

The results of the team on-going research in 2019 reflect this focus. Results presented in Sections 7.11 , 7.15 , 7.10 and 7.9 correspond to efforts for higher abstractions through DSL or libraries, and decouple algorithmics from parallel optimizations. Results in Section 7.8 correspond to efforts on parallelism expression and again abstraction, starting from standard parallel programming languages. Results described in Sections 7.1 and 7.16 provide feed-back information, through visualization and deadlock detection for parallel executions. The work described in Sections 7.3 , 7.4 , 7.5 , 7.6 ,7.12 , 7.7 and 7.13 focus in particular on StarPU and its development in order to better abstract architecture, resilience and optimizations. The work presented Section 7.2 aims to help developers with optimization.

Finally, Sections 7.14 and 7.17 present an on-going effort on improving the Chameleon library and strengthening its relation with StarPU and the NewMadeleine communication library. They represent real-life applications for the runtime methods we develop.

TADAAM Project-Team

3. Research Program

3.1. Need for System-Scale Optimization

Firstly, in order for applications to make the best possible use of the available resources, it is impossible to expose all the low-level details of the hardware to the program, as it would make impossible to achieve portability. Hence, the standard approach is to add intermediate layers (programming models, libraries, compilers, runtime systems, etc.) to the software stack so as to bridge the gap between the application and the hardware. With this approach, optimizing the application requires to express its parallelism (within the imposed programming model), organize the code, schedule and load-balance the computations, etc. In other words, in this approach, the way the code is written and the way it is executed and interpreted by the lower layers drives the optimization. In any case, this approach is centered on how computations are performed. Such an approach is therefore no longer sufficient, as the way an application is executing does depend less and less on the organization of computation and more and more on the way its data is managed.

Secondly, modern large-scale parallel platforms comprise tens to hundreds of thousand nodes [0]. However, very few applications use the whole machine. In general, an application runs only on a subset of the nodes [0]. Therefore, most of the time, an application shares the network, the storage and other resources with other applications running concurrently during its execution. Depending on the allocated resources, it is not uncommon that the execution of one application interferes with the execution of a neighboring one.

Lastly, even if an application is running alone, each element of the software stack often performs its own optimization independently. For instance, when considering an hybrid MPI/OpenMP application, one may realize that threads are concurrently used within the OpenMP runtime system, within the MPI library for communication progression, and possibly within the computation library (BLAS) and even within the application itself (pthreads). However, none of these different classes of threads are aware of the existence of the others. Consequently, the way they are executed, scheduled, prioritized does not depend on their relative roles, their locations in the software stack nor on the state of the application.

The above remarks show that in order to go beyond the state-of-the-art, it is necessary to design a new set of mechanisms allowing cross-layer and system-wide optimizations so as to optimize the way data is allocated, accessed and transferred by the application.

3.2. Scientific Challenges and Research Issues

In TADAAM, we will tackle the problem of efficiently executing an application, at system-scale, on an HPC machine. We assume that the application is already optimized (efficient data layout, use of effective libraries, usage of state-of-the-art compilation techniques, etc.). Nevertheless, even a statically optimized application will not be able to be executed at scale without considering the following dynamic constraints: machine topology, allocated resources, data movement and contention, other running applications, access to storage, etc. Thanks to the proposed layer, we will provide a simple and efficient way for already existing applications, as well as new ones, to express their needs in terms of resource usage, locality and topology, using a high-level semantic.

It is important to note that we target the optimization of each application independently but also several applications at the same time and at system-scale, taking into account their resource requirement, their network usage or their storage access. Furthermore, dealing with code-coupling application is an intermediate use-case that will also be considered.

[0] More than 22,500 XE6 compute node for the BlueWaters system; 5040 B510 Bullx Nodes for the Curie machine; more than 49,000 BGQ nodes for the MIRA machine.

[0] In 2014, the median case was 2048 nodes for the BlueWaters system and, for the first year of the Curie machine, the median case was 256 nodes

Several issues have to be considered. The first one consists in providing relevant **abstractions and models to describe the topology** of the available resources **and the application behavior**.

Therefore, the first question we want to answer is: **"How to build scalable models and efficient abstractions enabling to understand the impact of data movement, topology and locality on performance?"** These models must be sufficiently precise to grasp the reality, tractable enough to enable efficient solutions and algorithms, and simple enough to remain usable by non-hardware experts. We will work on (1) better describing the memory hierarchy, considering new memory technologies; (2) providing an integrated view of the nodes, the network and the storage; (3) exhibiting qualitative knowledge; (4) providing ways to express the multi-scale properties of the machine. Concerning abstractions, we will work on providing general concepts to be integrated at the application or programming model layers. The goal is to offer means, for the application, to express its high-level requirements in terms of data access, locality and communication, by providing abstractions on the notion of hierarchy, mesh, affinity, traffic metrics, etc.

In addition to the abstractions and the aforementioned models we need to **define a clean and expressive API in a scalable way**, in order for applications to express their needs (memory usage, affinity, network, storage access, model refinement, etc.).

Therefore, the second question we need to answer is: "**how to build a system-scale, stateful, shared layer that can gather applications needs expressed with a high-level semantic?**". This work will require not only to define a clean API where applications will express their needs, but also to define how such a layer will be shared across applications and will scale on future systems. The API will provide a simple yet effective way to express different needs such as: memory usage of a given portion of the code; start of a compute intensive part; phase where the network is accessed intensively; topology-aware affinity management; usage of storage (in read and/or write mode); change of the data layout after mesh refinement, etc. From an engineering point of view, the layer will have a hierarchical design matching the hardware hierarchy, so as to achieve scalability.

Once this has been done, the service layer, will have all the information about the environment characteristics and application requirements. We therefore need to design a set of **mechanisms to optimize applications execution**: communication, mapping, thread scheduling, data partitioning/mapping/movement, etc.

Hence, the last scientific question we will address is: "**How to design fast and efficient algorithms, mechanisms and tools to enable execution of applications at system-scale, in full a HPC ecosystem, taking into account topology and locality?**" A first set of research is related to thread and process placement according to the topology and the affinity. Another large field of study is related to data placement, allocation and partitioning: optimizing the way data is accessed and processed especially for mesh-based applications. The issues of transferring data across the network will also be tackled, thanks to the global knowledge we have on the application behavior and the data layout. Concerning the interaction with other applications, several directions will be tackled. Among these directions we will deal with matching process placement with resource allocation given by the batch scheduler or with the storage management: switching from a best-effort application centric strategy to global optimization scheme.

TRIBE Project-Team

3. Research Program

3.1. Research program

Following up on the effort initiated by the team members during the last few years and building on an approach combining protocol design, data analytics, and experimental research, we propose a research program organized around three closely related objectives that are briefly described in the following.

- **Technologies for accommodating low-end IoT devices:** The IoT is expected to gradually connect billions of low-end devices to the Internet, and thereby drastically increase communication without human source or destination. Low-end IoT devices differ starkly from high-end IoT devices in terms of resources such as energy, memory, and computational power. Projections show this divide will not fundamentally change in the future and that IoT should ultimately interconnect a dense population of devices as tiny as dust particles, feeding off ambient power sources (energy harvesting). These characteristics constrain the software and communication protocols running on low-end IoT devices: they are neither able to run a common software platform such as Linux (or its derivatives), nor the standard protocol stack based on TCP/IP. Solutions for low-end IoT devices require thus: **(i)optimized communication protocols** taking into account radio technology evolution and devices constrained requirements; **(ii)tailored software platforms** providing high level programming, modular software updates as well as advanced support for new security and energy concentration features; **(iii)unification of technologies** for low-end IoT, which is too fragmented at the moment, guaranteeing integration with core or other edge networks.
- **Technologies for leveraging high-end IoT devices' advents:** High-end IoT devices are one of the most important instances of the connected devices supporting a noteworthy shift towards mobile Internet access. As our lives become more dependent on pervasive connectivity, our social patterns (as human being in the Internet era) are nowadays being reflected from our real life onto the virtual binary world. This gives birth to two tendencies. From one side, edge networks can now be utilized as mirrors to reflect the inherent human dynamics, their context, and interests thanks to their well organized recording and almost ubiquitous coverage. From the other side, social norms and structure dictating human behavior (e.g., interactions, mobility, interest, cultural patterns) are now directly influencing the way individuals interact with the network services and demand resources or content. In particular, we observe the particularities present in human dynamics *shape the way (i.e., where, when, how, or what) resources, services, and infrastructures are used at the edge of the Internet*. Hence, we claim a need to digitally study high-end IoT devices' end-users behaviors and to leverage this understanding in networking solutions' design, so as to optimize network exploitation. This suggests the **integration of the heterogeneity and uncertainty of behaviors in designed networking solutions**. For this, *useful knowledge* allowing the understanding of behaviors and context of users has to be *extracted and delivered out* of large masses of data. Such knowledge has to be then *integrated in current design practices*. This brings the idea of a more *tactful networking design practice* where the network is assigned with the human like capability of observation, interpretation, and reaction to daily life features and entities involving high-end IoT devices. Research activities here include: **(i) the quest for meaningful data**, which includes the integration of data from different sources, the need for scaling up data analysis, the usage and analysis of fine-grained datasets, or still, the completion of sparse and coarse grained datasets; **(ii) expanding edge networks' usage understanding**, which concerns analysis on how and when contextual information impact network usage, fine-grained analysis of short-term mobility of individuals, or the identification of patterns of behavior and novelty-seeking of individuals; **(iii) human-driven prediction models**, extensible to context awareness and adapted to individuals preferences in terms of novelty, difersity, or routines.

- **Articulating the IoT edge with the core of the network:** The edge is the interface between the IoT devices and the core network: some of the challenges encountered by IoT devices have their continuity at the edge of the network inside the gateway (i.e., interoperability, heterogeneity and mobility support). Besides, the edge should be able to support intermediary functions between devices and the rest of the core (e.g., the cloud). This includes: **(i) proxying functionality**, facilitating connections between devices and the Internet; **(ii) machine learning enhanced IoT solutions**, designed to improve performance of advanced IoT networked systems (e.g., through methods such as supervised, unsupervised or reinforcement learning) at adapted levels of the protocol stack (e.g., for multiple access, coding, choices); **(iii) IoT data contextualization**, so collection of meaningful IoT data (i.e., right data collected at the right time) can be earlier determined closer to the data source; **(iv) intermediary computation** through fog or Mobile Edge Computing (MEC) models, where IoT devices can obtain computing, data storage, and communication means with lower latency in a decentralized way; or **(v) security of end-to-end IoT software supply-chain**, including remote management and over-the-air updates.

WHISPER Project-Team

3. Research Program

3.1. Scientific Foundations

3.1.1. Program analysis

A fundamental goal of the research in the Whisper team is to elicit and exploit the knowledge found in existing code. To do this in a way that scales to a large code base, systematic methods are needed to infer code properties. We may build on either static [36], [38], [40] or dynamic analysis [58], [60], [65]. Static analysis consists of approximating the behavior of the source code from the source code alone, while dynamic analysis draws conclusions from observations of sample executions, typically of test cases. While dynamic analysis can be more accurate, because it has access to information about actual program behavior, obtaining adequate test cases is difficult. This difficulty is compounded for infrastructure software, where many, often obscure, cases must be handled, and external effects such as timing can have a significant impact. Thus, we expect to primarily use static analyses. Static analyses come in a range of flavors, varying in the extent to which the analysis is *sound*, *i.e.*, the extent to which the results are guaranteed to reflect possible run-time behaviors.

One form of sound static analysis is *abstract interpretation* [38]. In abstract interpretation, atomic terms are interpreted as sound abstractions of their values, and operators are interpreted as functions that soundly manipulate these abstract values. The analysis is then performed by interpreting the program in a compositional manner using these abstracted values and operators. Alternatively, *dataflow analysis* [49] iteratively infers connections between variable definitions and uses, in terms of local transition rules that describe how various kinds of program constructs may impact variable values. Schmidt has explored the relationship between abstract interpretation and dataflow analysis [73]. More recently, more general forms of symbolic execution [36] have emerged as a means of understanding complex code. In symbolic execution, concrete values are used when available, and these are complemented by constraints that are inferred from terms for which only partial information is available. Reasoning about these constraints is then used to prune infeasible paths, and obtain more precise results. A number of works apply symbolic execution to operating systems code [33], [34].

While sound approaches are guaranteed to give correct results, they typically do not scale to the very diverse code bases that are prevalent in infrastructure software. An important insight of Engler et al. [42] was that valuable information could be obtained even when sacrificing soundness, and that sacrificing soundness could make it possible to treat software at the scales of the kernels of the Linux or BSD operating systems. Indeed, for certain types of problems, on certain code bases, that may mostly follow certain coding conventions, it may mostly be safe to *e.g.*, ignore the effects of aliases, assume that variable values are unchanged by calls to unanalyzed functions, etc. Real code has to be understood by developers and thus cannot be too complicated, so such simplifying assumptions are likely to hold in practice. Nevertheless, approaches that sacrifice soundness also require the user to manually validate the results. Still, it is likely to be much more efficient for the user to perform a potentially complex manual analysis in a specific case, rather than to implement all possible required analyses and apply them everywhere in the code base. A refinement of unsound analysis is the CEGAR approach [37], in which a highly approximate analysis is complemented by a sound analysis that checks the individual reports of the approximate analysis, and then any errors in reasoning detected by the sound analysis are used to refine the approximate analysis. The CEGAR approach has been applied effectively on device driver code in tools developed at Microsoft [25]. The environment in which the driver executes, however, is still represented by possibly unsound approximations.

Going further in the direction of sacrificing soundness for scalability, the software engineering community has recently explored a number of approaches to code understanding based on techniques developed in the areas of natural language understanding, data mining, and information retrieval. These approaches view code, as well as other software-reated artifacts, such as documentation and postings on mailing lists, as bags of words structured in various ways. Statistical methods are then used to collect words or phrases that seem to be highly correlated, independently of the semantics of the program constructs that connect them. The obliviousness to program semantics can lead to many false positives (invalid conclusions) [55], but can also highlight trends that are not apparent at the low level of individual program statements. We have previously explored combining such statistical methods with more traditional static analysis in identifying faults in the usage of constants in Linux kernel code [53].

3.1.2. Domain Specific Languages

Writing low-level infrastructure code is tedious and difficult, and verifying it is even more so. To produce non-trivial programs, we could benefit from moving up the abstraction stack to enable both programming and proving as quickly as possible. Domain-specific languages (DSLs), also known as *little languages*, are a means to that end [6] [61].

3.1.2.1. Traditional approach.

Using little languages to aid in software development is a tried-and-trusted technique [75] by which programmers can express high-level ideas about the system at hand and avoid writing large quantities of formulaic C boilerplate.

This approach is typified by the Devil language for hardware access [8]. An OS programmer describes the register set of a hardware device in the high-level Devil language, which is then compiled into a library providing C functions to read and write values from the device registers. In doing so, Devil frees the programmer from having to write extensive bit-manipulation macros or inline functions to map between the values the OS code deals with, and the bit-representation used by the hardware: Devil generates code to do this automatically.

However, DSLs are not restricted to being "stub" compilers from declarative specifications. The Bossa language [7] is a prime example of a DSL involving imperative code (syntactically close to C) while offering a high-level of abstraction. This design of Bossa enables the developer to implement new process scheduling policies at a level of abstraction tailored to the application domain.

Conceptually, a DSL both abstracts away low-level details and justifies the abstraction by its semantics. In principle, it reduces development time by allowing the programmer to focus on high-level abstractions. The programmer needs to write less code, in a language with syntax and type checks adapted to the problem at hand, thus reducing the likelihood of errors.

3.1.2.2. Embedding DSLs.

The idea of a DSL has yet to realize its full potential in the OS community. Indeed, with the notable exception of interface definition languages for remote procedure call (RPC) stubs, most OS code is still written in a low-level language, such as C. Where DSL code generators are used in an OS, they tend to be extremely simple in both syntax and semantics. We conjecture that the effort to implement a given DSL usually outweighs its benefit. We identify several serious obstacles to using DSLs to build a modern OS: specifying what the generated code will look like, evolving the DSL over time, debugging generated code, implementing a bug-free code generator, and testing the DSL compiler.

Filet-o-Fish (FoF) [39] addresses these issues by providing a framework in which to build correct code generators from semantic specifications. This framework is presented as a Haskell library, enabling DSL writers to *embed* their languages within Haskell. DSL compilers built using FoF are quick to write, simple, and compact, but encode rigorous semantics for the generated code. They allow formal proofs of the run-time behavior of generated code, and automated testing of the code generator based on randomized inputs, providing greater test coverage than is usually feasible in a DSL. The use of FoF results in DSL compilers that OS developers can quickly implement and evolve, and that generate provably correct code. FoF has been used

to build a number of domain-specific languages used in Barrelfish, [26] an OS for heterogeneous multicore systems developed at ETH Zurich.

The development of an embedded DSL requires a few supporting abstractions in the host programming language. FoF was developed in the purely functional language Haskell, thus benefiting from the type class mechanism for overloading, a flexible parser offering convenient syntactic sugar, and purity enabling a more algebraic approach based on small, composable combinators. Object-oriented languages – such as Smalltalk [43] and its descendant Pharo [30] – or multi-paradigm languages – such as the Scala programming language [63] – also offer a wide range of mechanisms enabling the development of embedded DSLs. Perhaps suprisingly, a low-level imperative language – such as C – can also be extended so as to enable the development of embedded compilers [27].

3.1.2.3. Certifying DSLs.

Whilst automated and interactive software verification tools are progressively being applied to larger and larger programs, we have not yet reached the point where large-scale, legacy software – such as the Linux kernel – could formally be proved "correct". DSLs enable a pragmatic approach, by which one could realistically strengthen a large legacy software by first narrowing down its critical component(s) and then focus our verification efforts onto these components.

Dependently-typed languages, such as Coq or Idris, offer an ideal environment for embedding DSLs [35], [31] in a unified framework enabling verification. Dependent types support the type-safe embedding of object languages and Coq's mixfix notation system enables reasonably idiomatic domain-specific concrete syntax. Coq's powerful abstraction facilities provide a flexible framework in which to not only implement and verify a range of domain-specific compilers [39], but also to combine them, and reason about their combination.

Working with many DSLs optimizes the "horizontal" compositionality of systems, and favors reuse of building blocks, by contrast with the "vertical" composition of the traditional compiler pipeline, involving a stack of comparatively large intermediate languages that are harder to reuse the higher one goes. The idea of building compilers from reusable building blocks is a common one, of course. But the interface contracts of such blocks tend to be complex, so combinations are hard to get right. We believe that being able to write and verify formal specifications for the pieces will make it possible to know when components can be combined, and should help in designing good interfaces.

Furthermore, the fact that Coq is also a system for formalizing mathematics enables one to establish a close, formal connection between embedded DSLs and non-trivial domain-specific models. The possibility of developing software in a truly "model-driven" way is an exciting one. Following this methodology, we have implemented a certified compiler from regular expressions to x86 machine code [48]. Interestingly, our development crucially relied on an existing Coq formalization, due to Braibant and Pous, [32] of the theory of Kleene algebras.

While these individual experiments seem to converge toward embedding domain-specific languages in rich type theories, further experimental validation is required. Indeed, Barrelfish is an extremely small software compared to the Linux kernel. The challenge lies in scaling this methodology up to large software systems. Doing so calls for a unified platform enabling the development of a myriad of DSLs, supporting code reuse across DSLs as well as providing support for mechanically-verified proofs.

3.2. Research direction: Tools for improving legacy infrastructure software

A cornerstone of our work on legacy infrastructure software is the Coccinelle program matching and transformation tool for C code. Coccinelle has been in continuous development since 2005. Today, Coccinelle is extensively used in the context of Linux kernel development, as well as in the development of other software, such as wine, python, kvm, and systemd. Currently, Coccinelle is a mature software project, and no research is being conducted on Coccinelle itself. Instead, we leverage Coccinelle in other research projects [28], [29], [64], [66], [70], [72], [74], [59], [54], both for code exploration, to better understand at a large scale problems in Linux development, and as an essential component in tools that require program matching and transformation. The continuing development and use of Coccinelle is also a source of visibility in the Linux kernel developer

community. We submitted the first patches to the Linux kernel based on Coccinelle in 2007. Since then, over 5500 patches have been accepted into the Linux kernel based on the use of Coccinelle, including around 3000 by over 500 developers from outside our research group.

Our recent work has focused on driver porting. Specifically, we have considered the problem of porting a Linux device driver across versions, particularly backporting, in which a modern driver needs to be used by a client who, typically for reasons of stability, is not able to update their Linux kernel to the most recent version. When multiple drivers need to be backported, they typically need many common changes, suggesting that Coccinelle could be applicable. Using Coccinelle, however, requires writing backporting transformation rules. In order to more fully automate the backporting (or symmetrically forward porting) process, these rules should be generated automatically. We have carried out a preliminary study in this direction with David Lo of Singapore Management University; this work, published at ICSME 2016 [77], is limited to a port from one version to the next one, in the case where the amount of change required is limited to a single line of code. Whisper has been awarded an ANR PRCI grant to collaborate with the group of David Lo on scaling up the rule inference process and proposing a fully automatic porting solution.

3.3. Research direction: developing infrastructure software using Domain Specific Languages

We wish to pursue a *declarative* approach to developing infrastructure software. Indeed, there exists a significant gap between the high-level objectives of these systems and their implementation in low-level, imperative programming languages. To bridge that gap, we propose an approach based on domain-specific languages (DSLs). By abstracting away boilerplate code, DSLs increase the productivity of systems programmers. By providing a more declarative language, DSLs reduce the complexity of code, thus the likelihood of bugs.

Traditionally, systems are built by accretion of several, independent DSLs. For example, one might use Devil [8] to interact with devices, Bossa [7] to implement the scheduling policies. However, much effort is duplicated in implementing the back-ends of the individual DSLs. Our long term goal is to design a unified framework for developing and composing DSLs, following our work on Filet-o-Fish [39]. By providing a single conceptual framework, we hope to amortize the development cost of a myriad of DSLs through a principled approach to reusing and composing them.

Beyond the software engineering aspects, a unified platform brings us closer to the implementation of mechanically-verified DSLs. Using the Coq proof assistant as an x86 macro-assembler [48] is a step in that direction, which belongs to a larger trend of hosting DSLs in dependent type theories [31], [35], [62]. A key benefit of those approaches is to provide – by construction – a formal, mechanized semantics to the DSLs thus developed. This semantics offers a foundation on which to base further verification efforts, whilst allowing interaction with non-verified code. We advocate a methodology based on incremental, piece-wise verification. Whilst building fully-certified systems from the top-down is a worthwhile endeavor [50], we wish to explore a bottom-up approach by which one focuses first and foremost on crucial subsystems and their associated properties.

Our current work on DSLs has two complementary goals: (i) the design of a unified framework for developing and composing DSLs, following our work on Filet-o-Fish, and (ii) the design of domain-specific languages for domains where there is a critical need for code correctness, and corresponding methodologies for proving properties of the run-time behavior of the system.

WIDE Project-Team

3. Research Program

3.1. Overview

In order to progress in the four fields described above, the WIDE team is developing a research program which aims to **help developers control and master the inherent uncertainties and performance challenges brought by scale and distribution**.

More specifically, our program revolves around four key challenges.

- Challenge 1: Designing Hybrid Scalable Architectures,
- Challenge 2: Constructing Personalizable Privacy-Aware Distributed Systems,
- Challenge 3: Understanding Controllable Network Diffusion Processes,
- Challenge 4: Systemizing Modular Distributed Computability and Efficiency.

These four challenges have in common **the inherent tension between coordination and scalability in large-scale distributed systems**: strong coordination mechanisms can deliver strong guarantees (in terms of consistency, agreement, fault-tolerance, and privacy protection), but are generally extremely costly and inherently non-scalable if applied indiscriminately. By contrast, highly scalable coordination approaches (such as epidemic protocols, eventual consistency, or self-organizing overlays) perform much better when the size of a system increases, but do not, in most cases, provide any strong guarantees in terms of consistency or agreement.

The above four challenges explore these tensions from *four complementary angles*: from an architectural perspective (Challenge 1), from the point of view of a fundamental system-wide guarantee (privacy protection, Challenge 2), looking at one universal scalable mechanism (network diffusion, Challenge 3), and considering the interplay between modularity and computability in large-scale systems (Challenge 4). These four challenges range from practical concerns (Challenges 1 and 2) to more theoretical questions (Challenrges 3 and 4), yet present *strong synergies* and *fertile interaction points*. E.g. better understanding network diffusion (Challenge 3) is a key enabler to develop more private decentralized systems (Challenge 2), while the development of a theoretically sound modular computability hierarchy (Challenge 4) has a direct impact on our work on hybrid architectures (Challenge 1).

3.2. Hybrid Scalable Architectures

The rise of planetary-scale distributed systems calls for novel software and system architectures that can support user-facing applications while scaling to large numbers of devices, and leveraging established and emerging technologies. The members of WIDE are particularly well positioned to explore this avenue of research thanks to their experience on de-concentrated architectures combining principles from both decentralized peer-to-peer [48], [58] systems and hybrid infrastructures (i.e. architectures that combines centralized or hierarchical elements, often hosted in well-provisioned data-centers, and a decentralized part, often hosted in a peer-to-peer overlay) [52]. In the short term, we aim to explore two axes in this direction: browser-based communication, and micro services.

3.2.1. Browser-based fog computing

The dramatic increase in the amount of data being produced and processed by connected devices has led to paradigms that seek to decentralize the traditional cloud model. In 2011 Cisco [49] introduced the vision of *fog computing* that combines the cloud with resources located at the edge of the network and in between. More generally, the term *edge computing* has been associated with the idea of adding edge-of-the network storage and computation to traditional cloud infrastructures [44].

A number of efforts in this directions focus on specific hardware, e.g. fog nodes that are responsible for connected IoT devices [50]. However, many of today's applications run within web browsers or mobile phones. In this context, the recent introduction of the WebRTC API, makes it possible for browsers and smartphones to exchange directly between each other, enabling mobile, or browser-based decentralized applications. Maygh [72], for example, uses the WebRTC API to build a decentralized Content Delivery Network that runs solely on web browsers. The fact that the application is hosted completely on a web server and downloaded with enabled websites means that webmasters can adopt the Content Delivery Network (CDN) without requiring users to install any specific software.

For us, the ability of browsers to communicate with each other using the WebRTC paradigm provides a novel playground for new programming models, and for a *browser-based fog architecture* combining both a centralized, cloud-based part, and a decentralized, browser-supported part.

This model offers tremendous potential by making edge-of-the-network resources available through the interconnection of web-browsers, and offers new opportunities for the protection of the personal data of end users. But consistently engineering browser-based components requires novel tools and methodologies.

In particular, WebRTC was primarily designed for exchanging media and data between two browsers in the presence of a coordinating server. Its complex mechanisms for connection establishment make many of the existing peer-to-peer protocols inefficient. To address this challenge, we plan to consider two angles of attack. First, we plan to design novel protocols that take into account the specific requirements set by this new technology. Second, we envisage to investigate variants of the current WebRTC model with cheaper connection-establishment protocols, in order to provide lower delays and bandwidth consumption in large-scale browser-based applications.

We also plan to address the trade-offs associated with hybrid browser-cloud models. For example, when should computation be delegated to browsers and when should it be executed on the cloud in order to maximize the quality of service? Or, how can a decentralized analytics algorithms operating on browser-based data complement or exploit the knowledge built by cloud-based data analytics solutions?

3.2.2. Emergent micro-service deployment and management

Micro-services tend to produce fine-grained applications in which many small services interact in a loosely coupled manner to produce a wide range of services within an organization. Individual services need to evolve independently of each other over time without compromising the availability of the overall application. Lightweight isolation solutions such as containers (Docker, ...), and their associated tooling ecosystem (e.g. Google's Borg [71], Kubernetes [47]) have emerged to facilitate the deployment of large-scale micro-service-based applications, but only provide preliminary solutions for key concerns in these systems, which we would like to investigate and extend.

Most of today's on-line computer systems are now too large to evolve in monolithic, entirely pre-planned ways. This applies to very large data centres, for example, where the placement of virtual machines to reduce heating and power consumption can no longer be treated using top-down exhaustive optimisation approaches beyond a critical size. This is also true of social networking applications, where different mechanisms—e.g. to spread news notifications, or to recommend new contacts—must be adapted to the different sub-communities present in the system.

To cope with the inherent complexity of building complex loosely-coupled distributed systems while fostering and increasing efficiency, maintainability, and scalability, we plan to study how novel programming techniques based on declarative programming, components and epidemic protocols can help design, deploy, and maintain self-adaptive structures (e.g. placement of VM) and mechanisms (e.g. contact recommendations) that are optimized to the local context of very large distributed systems. To fulfill this vision, we plan to explore a three-pronged strategy to raise the level of programming abstraction offered to developers.

- First, we plan to explore the use of high-level domain-specific languages (DSL) to declare how large-scale topologies should be achieved, deployed, and maintained. Our vision is a declarative approach to describe how to combine, deploy and orchestrate micro-services in an abstract manner

thus abstracting away developers from the underlying cloud infrastructures, and from the intricacies involved in writing low-level code to build a large-scale distributed application that scales. With this effort, we plan notably to directly support the twin properties of *emergence* (the adaptation "from within") and *differentiation* (the possibility from parts of the system to diverge while still forming a whole). Our central objective is to search for principled programming constructs to support these two capabilities using a modular and incremental software development approach.

- On a second strand of work, we plan to investigate how unikernels enable smaller footprints, more optimization options, and faster boot times for micro-services. Isolating micro-services into VMs is not the most adequate approach as it requires the use of hypervisors, or virtual machine monitors (VMMs), to virtualize hardware resources. VMMs are well known to be heavyweight with both boot and run time overheads that may have a strong impact on performances. Unikernels seem to offer the right balance between performance and flexibility to address this challenge. One of the key underlying challenges is to compile directly the aforementioned provided DSL to a dedicated and customized machine image, ready to be deployed directly on top of a large set of bare metal servers.
- Depending on the workload it is subjected to, and the state of its execution environment (network, VMs), a large-scale distributed application may present erratic or degraded performance that is hard to anticipate and plan for. There is therefore a strong need to adapt dynamically the way resources are allocated to a running application. We would like to study how the DSL approach we envisage can be extended to enable developers to express orchestration algorithms based on machine learning algorithms.

3.3. Personalizable Privacy-Aware Distributed Systems

On-line services are increasingly moving towards an in-depth analysis of user data, with the objective of providing ever better personalization. But in doing so, personalized on-line services inevitably pose risks to the privacy of users. Eliminating, or even reducing these risks raises important challenges caused by the inherent trade-off between the level of personalization users wish to achieve, and the amount of information they are willing to reveal about themselves (explicitly or through the many implicit sources of digital information such as smart homes, smart cars, and IoT environments).

At a general level, we would like to address these challenges through protocols that can provide access to unprecedented amounts of data coming from sensors, users, and documents published by users, while protecting the privacy of individuals and data sources. To this end, we plan to rely on our experience in the context of distributed systems, recommender systems, and privacy, as well as in our collaborations with experts in neighboring fields such as machine learning, and security. In particular, we aim to explore different privacy-utility tradeoffs that make it possible to provide differentiated levels of privacy guarantees depending on the context associated with data, on the users that provide the data, and on those that access it. Our research targets the general goal of privacy-preserving decentralized learning, with applications in different contexts such as user-oriented applications, and the Internet-of-Things (IoT).

3.3.1. Privacy-preserving decentralized learning

Personalization and recommendation can be seen as a specific case of general machine learning. Production-grade recommenders and personalizers typically centralize and process the available data in one location (a data-center, a cloud service). This is highly problematic, as it endangers the privacy of users, while hampering the analysis of datasets subject to privacy constraints that are held by multiple independent organizations (such as health records). A decentralized approach to machine learning appears as a promising candidate to overcome these weaknesses: if each user or participating organization keeps its data, while only exchanging gradient or model information, privacy leaks seem less likely to occur.

In some cases, decentralized learning may be achieved through relatively simple adaptations of existing centralized models, for instance by defining alternative learning models that may be more easily decentralized. But in all cases, processing growing amounts of information calls for high-performance algorithms and middleware that can handle diverse storage and computation resources, in the presence of dynamic and

privacy-sensitive data. To reach this objective, we will therefore leverage our work in distributed and privacy-preserving algorithms and middleware [51], [53], [54] as well as the results of our work on large-scale hybrid architectures in Objective 1.

3.3.2. Personalization in user-oriented applications

As a first application perspective, we plan to design tools that exploit decentralized analytics to enhance user-centric personalized applications. As we observed above, such applications exhibit an inherent trade-off between personalization quality and privacy preservation. The most obvious goal in this direction consists in designing algorithms that can achieve high levels of personalization while protecting sensitive user information. But an equally important one consists in personalizing the trade-off itself by adapting the quality of the personalization provided to a user to his/her willingness to expose information. This, like other desirable behaviors, appears at odds with the way current systems work. For example, a user of a recommender system that does not reveal his/her profile information penalizes other users causing them to receive less accurate recommendations. We would like to mitigate this situation by means of protocols that reward users for sharing information. On the one hand, we plan to take inspiration from protocols for free-riding avoidance in peer-to-peer systems [55], [60]. On the other hand, we will consider blockchains as a tool for tracking and rewarding data contributions. Ultimately, we aim at enabling users to configure the level of privacy and personalization they wish to experience.

3.3.3. Privacy preserving decentralized aggregation

As a second setting we would like to consider target applications running on constrained devices like in the Internet-of-Things (IoT). This setting makes it particularly important to operate on decentralized data in a light-weight privacy-preserving manner, and further highlights the synergy between this objective and Objective 1. For example, we plan to provide data subjects with the possibility to store and manage their data locally on their own devices, without having to rely on third-party managers or aggregators, but possibly storing less private information or results in the cloud. Using this strategy, we intend to design protocols that enable users themselves, or third-party companies to query distributed data in aggregate form, or to run data analytics processes on a distributed set of data repositories, thereby gathering knowledge without violating the privacy of other users. For example, we have started working on the problem of computing an aggregate function over a subset of the data in a distributed setting. This involves two major steps: selection and aggregation. With respect to selection, we envision defining a decentralized data-selection operation that can apply a selection predicate without violating privacy constraints. With respect to aggregation, we will continue our investigation of lightweight protocols that can provide privacy with limited computational complexity [45].

3.4. Network Diffusion Processes

Social, biological, and technological networks can serve as conduits for the spread of ideas, trends, diseases, or viruses. In social networks, rumors, trends and behaviors, or the adoption of new products, spread from person to person. In biological networks, diseases spread through contact between individuals, and mutations spread from an individual to its offsprings. In technological networks, such as the Internet and the power grid, viruses and worms spread from computer to computer, and power failures often lead to cascading failures. The common theme in all the examples above is that the rumor, disease, or failure starts out with a single or a few individual nodes, and propagates through the network, from node to node, to reach a potentially much larger number of nodes.

These types of *network diffusion processes* have long been a topic of study in various disciplines, including sociology, biology, physics, mathematics, and more recently, computer science. A main goal has been to devise mathematical models for these processes, describing how the state of an individual node can change as a function of the state of its neighbors in the network, and then analyse the role of the network structure in the outcome of the process. Based on our previous work, we would like to study to what extent one can affect the outcome of the diffusion process by controlling a small, possibly carefully selected fraction of the network.

For example, we plan to explore how we may increase the spread or speed of diffusion by choosing an appropriate set of seed nodes (a standard goal in viral marketing by word-of-mouth), or achieve the opposite effect either by choosing a small set of nodes to remove (a goal in immunization against diseases), or by seeding a competing diffusion (e.g., to limit the spread of misinformation in a social network).

Our goal is to provide a framework for a systematic and rigorous study of these problems. We will consider several standard diffusion models and extensions of them, including models from mathematical sociology, mathematical epidemiology, and interacting particle systems. We will consider existing and new variants of spread maximization/limitation problems, and will provide (approximation) algorithms or show negative (inapproximability) results. In case of negative results, we will investigate general conditions that make the problem tractable. We will consider both general network topologies and specific network models, and will relate the efficiency of solutions to structural properties of the topology. Finally, we will use these insights to engineer new network diffusion processes for efficient data dissemination.

3.4.1. Spread maximization

Our goal is in particular to study spread maximization in a broader class of diffusion processes than the basic independent cascade (IC) and linear threshold (LT) models of influence [64], [65], [66] that have been studied in this context so far. This includes the *randomized rumor spreading (RS)* model for information dissemination [57], *biased* versions of the *voter model* [61] modelling influence, and the (graph-based) *Moran processes* [68] modelling the spread of mutations. We would like to consider several natural versions of the spread maximization problem, and the relationships between them. For these problems we will use the greedy algorithm and the submodularity-based analytical framework of [64], and will also explore new approaches.

3.4.2. Immunization optimization

Conversely we would also like to explore immunization optimization problems. Existing works on these types of problem assume a *perfect-contagion* model, i.e., once a node gets infected, it deterministically infects all its non-immunized neighbors. We plan to consider various diffusion processes, including the standard *susceptible–infected* (SI), *susceptible–infected–recovered* (SIR) and *susceptible–infected–susceptible* (SIS) epidemic models, and explore the extent to which results and techniques for the perfect-contagion model carry over to these probabilistic models. We will also investigate whether techniques for spread maximization could be applied to immunization problems.

Some immunization problems are known to be hard to approximate in general graphs, even for the perfect-contagion model, e.g., the fixed-budget version of the fire-fighter problem cannot be approximated to any $n^{1-\epsilon}$ factor [46]. This strand of work will consider restricted graph families, such as trees or graphs of small treewidth, for such problems. In addition, for some immunization problems, there is a large gap between the best known approximation algorithm and the best known inaproximability result, and we would like to make progress in reducing these gaps.

3.5. Systemizing Modular Distributed Computability and Efficiency

The applications and services envisaged in Objectives 1 and 2 will lead to increasingly complex and multifaceted systems. Constructing these novel hybrid and decentralized systems will naturally push our need to understand distributed computing beyond the current state of the art. These trends therefore demand research efforts in establishing sound theoretical foundations to allow everyday developers to master the design, properties and implementation of these systems. We plan to investigate these foundations along two directions: first by studying novel approaches to some fundamental problems of *mutual exclusion and distributed coordination*, and second by exploring how we can build a *comprehensive and modular framework* capturing the foundations of *distributed computation*.

3.5.1. Randomized algorithm for mutual exclusion and coordination

To exploit the power of massive distributed applications and systems (such as those envisaged in Objectives 1 and 2) or multiple processors, algorithms must cope with the scale and asynchrony of these systems, and their inherent instability, e.g., due to node, link, or processor failures. Our goal is to explore the power and limits of randomized algorithms for large-scale networks of distributed systems, and for shared memory multiprocessor systems, in effect providing fundamental building blocks to the work envisioned in Objectives 1 and 2.

For shared memory systems, randomized algorithms have notably proved extremely useful to deal with asynchrony and failures. Sometimes probabilistic algorithms provide the only solution to a problem; sometimes they are more efficient; sometimes they are simply easier to implement. We plan to devise efficient algorithms for some of the fundamental problems of shared memory computing, such as mutual exclusion, renaming, and consensus.

In particular, looking at the problem of *mutual exclusion*, it is desirable that mutual exclusion algorithms be *abortable*. This means that a process that is trying to lock the resource can abort its attempt in case it has to wait too long. Abortability is difficult to achieve for mutual exclusion algorithms. We will try to extend our algorithms for the *cache-coherent* (CC) and the *distributed shared memory* (DSM) model in order to make them abortable, while maintaining expected constant *Remote Memory References* (RMRs) complexity, under optimistic system assumptions. In order to achieve this, the algorithm will use strong synchronization primitives, called compare-and-swap objects. As part of our collaboration with the University of Calgary, we will work on implementing those objects from registers in such a way that they also allow aborts. Our goal is to build on existing non-abortable implementations [59]. We plan then later to use these objects as building blocks in our mutual exclusion algorithm, in order to make them work even if the system does not readily provide such primitives.

We have also started working on blockchains, as these represent a new and interesting trade-off between probabilistic guarantees, scalability, and system dynamics, while revisiting some of the fundamental questions and limitations of consensus in fault-prone asynchronous systems.

3.5.2. Modular theory of distributed computing

Practitioners and engineers have proposed a number of reusable frameworks and services to implement specific distributed services (from Remote Procedure Calls with Java RMI or SOAP-RPC, to JGroups for group communication, and Apache Zookeeper for state machine replication). In spite of the high conceptual and practical interest of such frameworks, many of these efforts lack a sound grounding in distributed computation theory (with the notable exceptions of JGroups and Zookeeper), and often provide punctual and partial solutions for a narrow range of services. We argue that this is because we still lack a generic framework that unifies the large body of fundamental knowledge on distributed computation that has been acquired over the last 40 years.

To overcome this gap we would like to develop a systematic model of distributed computation that organizes the functionalities of a distributed computing system into reusable modular constructs assembled via well-defined mechanisms that maintain sound theoretical guarantees on the resulting system. This research vision arises from the strong belief that distributed computing is now mature enough to resolve the tension between the social needs for distributed computing systems, and the lack of a fundamentally sound and systematic way to realize these systems.

To progress on this vision, we plan in the near future to investigate, from a distributed software point of view, the impact due to failures and asynchrony on the layered architecture of distributed computing systems. A first step in this direction will address the notions of *message adversaries* (introduced a long time ago in [70]) and *process adversaries* (investigated in several papers, e.g. [69], [56], [62], [63], [67]). The aim of these notions is to consider failures, not as "bad events", but as part of the normal behavior of a system. As an example, when considering round-based algorithms, a message adversary is a daemon which, at every round, is allowed to suppress some messages. The aim is then, given a problem P, to find the strongest adversary under which P

can be solved ("strongest" means here that giving more power to the adversary makes the problem impossible to solve). This work will allow us to progress in terms of general *layered* theory of distributed computing, and allow us to better *map* distributed computing models and their relations, in the steps of noticeable early efforts in this direction [69], [43].

ALICE Team

3. Research Program

3.1. Point clouds

Currently, transforming the raw point cloud into a triangular mesh is a long pipeline involving disparate geometry processing algorithms:

- *Point pre-processing:* colorization, filtering to remove unwanted background, first noise reduction along acquisition viewpoint;
- *Registration:* cloud-to-cloud alignment, filtering of remaining noise, registration refinement;
- *Mesh generation:* triangular mesh from the complete point cloud, re-meshing, smoothing.

The output of this pipeline is a locally structured model which is used in downstream mesh analysis methods such as feature extraction, segmentation in meaningful parts or building CAD models.

It is well known that point cloud data contains measurement errors due to factors related to the external environment and to the measurement system itself [37], [32], [20]. These errors propagate through all processing steps: pre-processing, registration and mesh generation. Even worse, the heterogeneous nature of different processing steps makes it extremely difficult to know *how* these errors propagate through the pipeline. To give an example, for cloud-to-cloud alignment it is necessary to estimate normals. However, the normals are forgotten in the point cloud produced by the registration stage. Later on, when triangulating the cloud, the normals are re-estimated on the modified data, thus introducing uncontrollable errors.

We plan to develop new reconstruction, meshing and re-meshing algorithms, with a specific focus on the accuracy and resistance to all defects present in the input raw data. We think that pervasive treatment of uncertainty is the missing ingredient to achieve this goal. We plan to rethink the pipeline with the position uncertainty maintained during the whole process. Input points can be considered either as error ellipsoids [41] or as probability measures [27]. In a nutshell, our idea is to start by computing an error ellipsoid [43], [29] for each point of the raw data, and then to cumulate the errors (approximations) committed at each step of the processing pipeline while building the mesh. In this way, the final users will be able to take the uncertainty knowledge into account and rely on this confidence measure for further analysis and simulations. Quantifying uncertainty for reconstruction algorithms, and propagating them from input data to high-level geometry processing algorithms has never been considered before, possibly due to the very different methodologies of the approaches involved. At the very beginning we will re-implement the entire pipeline, and then attack the weak links through all three reconstruction stages.

3.2. Parameterizations

One of the favorite tools we use in our team are parameterizations. They provide a very powerful way to reveal structures on objects. The most omnipresent application of parameterizations is texture mapping: texture maps provide a way to represent in 2D (on the map) information related to a surface. Once the surface is equipped with a map, we can do much more than a mere coloring of the surface: we can approximate geodesics, edit the mesh directly in 2D or transfer information from one mesh to another.

Parameterizations constitute a family of methods that involve optimizing an objective function, subject to a set of constraints (equality, inequality, being integer, etc.). Computing the exact solution to such problems is beyond any hope, therefore approximations are the only resort. This raises a number of problems, such as the minimization of highly nonlinear functions and the definition of direction fields topology, without forgetting the robustness of the software that puts all this into practice.

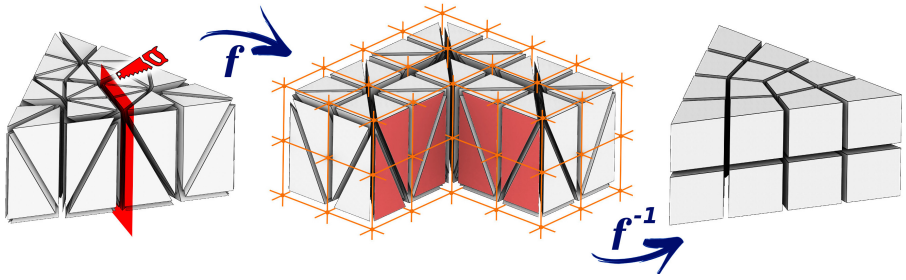


Figure 2. Hex-remeshing via global parameterization. ***Left:*** *Input tetrahedral mesh. To allow for a singular edge in the center, the mesh is cut open along the red plane.* ***Middle:*** *Mesh in parametric space.* ***Right:*** *Output mesh defined by parameterization.*

We are particularly interested in a specific instance of parameterization: hexahedral meshing. The idea [4] is to build a transformation f from the domain to a parametric space, where the deformed domain can be meshed by a regular grid. The inverse transformation f^{-1} applied to this grid produces the hexahedral mesh of the domain, aligned with the boundary of the object. The strength of this approach is that the transformation may admit some discontinuities. Let us show an example: we start from a tetrahedral mesh (Figure 2 , left) and we want deform it in a way that its boundary is aligned with the integer grid. To allow for a singular edge in the output (the valency 3 edge, Figure 2 , right), the input mesh is cut open along the highlighted faces and the central edge is mapped onto an integer grid line (Figure 2 , middle). The regular integer grid then induces the hexahedral mesh with the desired topology.

Current global parameterizations allow grids to be positioned inside geometrically simple objects whose internal structure (the singularity graph) can be relatively basic. We wish to be able to handle more configurations by improving three aspects of current methods:

- Local grid orientation is usually prescribed by minimizing the curvature of a 3D steering field. Unfortunately, this heuristic does not always provide singularity curves that can be integrated by the parameterization. We plan to explore how to embed integrability constraints in the generation of the direction fields. To address the problem, we already identified necessary validity criteria, for example, the permutation of axes along elementary cycles that go around a singularity must preserve one of the axes (the one tangent to the singularity). The first step to enforce this (necessary) condition will be to split the frame field generation into two parts: first we will define a locally stable vector field, followed by the definition of the other two axes by a 2.5D directional field (2D advected by the stable vector field).
- The grid combinatorial information is characterized by a set of integer coefficients whose values are currently determined through numerical optimization of a geometric criterion: the shape of the hexahedra must be as close as possible to the steering direction field. Thus, the number of layers of hexahedra between two surfaces is determined solely by the size of the hexahedra that one wishes to generate. In this setting degenerate configurations arise easily, and we want to avoid them. In practice, mixed integer solvers often choose to allocate a negative or zero number of layers of hexahedra between two constrained sheets (boundaries of the object, internal constraints or singularities). We will study how to inject strict positivity constraints into these cases, which is a very complex problem because of the subtle interplay between different degrees of freedom of the system. Our first results for quad-meshing of surfaces give promising leads, notably thanks to *motorcycle graphs* [21], a notion we wish to extend to volumes.
- Optimization for the geometric criterion makes it possible to control the average size of the hexahedra, but it does not ensure the bijectivity (even locally) of the resulting parameterizations.

Considering other criteria, as we did in 2D [26], would probably improve the robustness of the process. Our idea is to keep the geometry criterion to find the global topology, but try other criteria to improve the geometry.

3.3. Hexahedral-dominant meshing

All global parameterization approaches are decomposed into three steps: frame field generation, field integration to get a global parameterization, and final mesh extraction. Getting a full hexahedral mesh from a global parameterization means that it has positive Jacobian everywhere except on the frame field singularity graph. To our knowledge, there is no solution to ensure this property, but some efforts are done to limit the proportion of failure cases. An alternative is to produce hexahedral dominant meshes. Our position is in between those two points of view:

1. We want to produce full hexahedral meshes;
2. We consider as pragmatic to keep hexahedral dominant meshes as a fallback solution.

The global parameterization approach yields impressive results on some geometric objects, which is encouraging, but not yet sufficient for numerical analysis. Note that while we attack the remeshing with our parameterizations toolset, the wish to improve the tool itself (as described above) is orthogonal to the effort we put into making the results usable by the industry. To go further, our idea (as opposed to [30], [22]) is that the global parameterization should not handle all the remeshing, but merely act as a guide to fill a large proportion of the domain with a simple structure; it must cooperate with other remeshing bricks, especially if we want to take final application constraints into account.

For each application we will take as an input domains, sets of constraints and, eventually, fields (e.g. the magnetic field in a tokamak). Having established the criteria of mesh quality (per application!) we will incorporate this input into the mesh generation process, and then validate the mesh by a numerical simulation software.

ALMANACH Project-Team

3. Research Program

3.1. Research strands

As described above, ALMAnaCH's scientific programme is organised around three research axes. The first two aim to tackle the challenge of language variation in two complementary directions. They are supported by a third, transverse research axis on language resources. Our four-year objectives are described in much greater detail in the project-team proposal, whose very recent final validation in June 2019 resulted in the upgrade of ALMAnaCH to the "project-team" status in July 2019. They can be summarised as follows:

3.1.1. Research axis 1

Our first objective is to **stay at a state-of-the-art level in key NLP tasks** such as shallow processing, part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. This will also require us to improve the **generation of semantic representations (semantic parsing)**, and to begin to explore tasks such as machine translation, which now relies on neural architectures also used for some of the above-mentioned tasks. Given the generalisation of neural models in NLP, we will also be involved in better understanding how such models work and what they learn, something that is directly related to the investigation of language variation (Research axis 2). We will also work on the **integration of both linguistic and non-linguistic contextual information** to improve automatic linguistic analysis. This is an emerging and promising line of research in NLP. We will have to identify, model and take advantage of each type of contextual information available. Addressing these issues will enable the development of new lines of research related to conversational content. Applications include improved information and knowledge extraction algorithms. We will especially focus on challenging datasets such as domain-specific texts (e.g. financial, legal) as well as historical documents, in the larger context of the development of digital humanities. We currently also explore the even more challenging new direction of a cognitively inspired NLP, in order to tackle the possibility to enrich the architecture of state-of-the-art algorithms, such as RNNGs, based on human neuroimaging-driven data.

3.1.2. Research axis 2

Language variation must be better understood and modelled in all its forms. In this regard, we will put a strong emphasis on **four types** of language variation and their mutual interaction: **sociolinguistic variation** in synchrony (including non-canonical spelling and syntax in user-generated content), **complexity-based variation** in relation to language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus on Old to all forms of Modern French, as well as Indo-European languages in general). In addition, the noise introduced by Optical Character Recognition and Handwritten Text Recognition systems, especially in the context of historical documents, bears some similarities to that of non-canonical input in user-generated content (e.g. erroneous characters). This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. Other types of language variation will also become important research topics for ALMAnaCH in the future. This includes dialectal variation (e.g. work on Arabic varieties, something on which we have already started working, producing the first annotated data set on Maghrebi Arabizi, the Arabic variants used on social media by people from North-African countries, written using a non-fixed Latin-script transcription) as well as the study and exploitation of paraphrases in a broader context than the above-mentioned complexity-based variation.

Both research axes above rely on the availability of language resources (corpora, lexicons), which is the focus of our third, transverse research axis.

3.1.3. Research axis 3

Language resource development (raw and annotated corpora, lexical resources) is not just a necessary preliminary step to create both evaluation datasets for NLP systems and training datasets for NLP systems based on machine learning. When dealing with datasets of interest to researchers from the humanities (e.g. large archives), it is also a goal *per se* and a preliminary step before making such datasets available and exploitable online. It involves a number of scientific challenges, among which (i) tackling issues related to the digitalisation of non-electronic datasets, (ii) tackling issues related to the fact that many DH-related datasets are domain-specific and/or not written in contemporary languages; (iii) the development of semi-automatic and automatic algorithms to speed up the work (e.g. automatic extraction of lexical information, low-resource learning for the development of pre-annotation algorithms, transfer methods to leverage existing tools and/or resources for other languages, etc.) and (iv) the development of formal models to represent linguistic information in the best possible way, thus requiring expertise at least in NLP and in typological and formal linguistics. Such endeavours are domains of expertise of the ALMAnaCH team, and a large part of our research activities will be dedicated to language resource development. In this regard, we aim to retain our leading role in the representation and management of lexical resource and treebank development and also to develop a complete processing line for the transcription, analysis and processing of complex documents of interest to the humanities, in particular archival documents. This research axis 3 will benefit the whole team and beyond, and will benefit from and feed the work of the other research axes.

3.2. Automatic Context-augmented Linguistic Analysis

This first research strand is centred around NLP technologies and some of their applications in Artificial Intelligence (AI). Core NLP tasks such as part-of-speech tagging, syntactic and semantic parsing is improved by integrating new approaches, such as (deep) neural networks, whenever relevant, while preserving and taking advantage of our expertise on symbolic and statistical system: hybridisation not only couples symbolic and statistical approaches, but neural approaches as well. AI applications are twofold, notwithstanding the impact of language variation (see the next strand): (i) information and knowledge extraction, whatever the type of input text (from financial documents to ancient, historical texts and from Twitter data to Wikipedia) and (ii) chatbots and natural language generation. In many cases, our work on these AI applications is carried out in collaboration with industrial partners. The specificities and issues caused by language variation (a text in Old French, a contemporary financial document and tweets with a non-canonical spelling cannot be processed in the same way) are addressed in the next research strand.

3.2.1. Processing of natural language at all levels: morphology, syntax, semantics

Our expertise in NLP is the outcome of more than 10 years in developing new models of analysis and accurate techniques for the full processing of any kind of language input since the early days of the Atoll project-team and the rise of linguistically informed data-driven models as put forward within the Alpage project-team.

Traditionally, a full natural language process (NLP) chain is organised as a pipeline where each stage of analysis represents a traditional linguistic field (in a *structuralism* view) from morphological analysis to purely semantic representations. The problem is that this architecture is vulnerable to error propagation and very domain sensitive: each of these stage must be compatible at the lexical and structure levels they provide. We arguably built the best performing NLP chain for French [74], [112] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [108], [105], [111], [82]. So we pursue our efforts on each of our components we developed: tokenisers (e.g. SxPipe), part-of-speech taggers (e.g. MElt), constituency parsers and dependency parsers (e.g. FRMG, DyALog-SR) as well as our recent neural semantic graph parsers [105].

In particular, we continue to explore the hybridisation of symbolic and statistical approaches, and extend it to neural approaches, as initiated in the context of our participation to the CoNLL 2017 multilingual parsing shared task [0] and to Extrinsic Parsing Evaluation Shared Task [0].

[0] We ranked 3 for UPOS tagging and 6 for dependency parsing out of 33 participants.

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goal is to explore techniques such as multi-task learning (cf. already [110]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

3.2.2. Integrating context in NLP systems

While those points are fundamental, therefore necessary, if we want to build the next generation of NLP tools, we need to *push the envelop* even further by tackling the biggest current challenge in NLP: handling the context within which a speech act is taking place.

There is indeed a strong tendency in NLP to assume that each sentence is independent from its siblings sentences as well as its context of enunciation, with the obvious objective to simplify models and reduce the complexity of predictions. While this practice is already questionable when processing full-length edited documents, it becomes clearly problematic when dealing with short sentences that are noisy, full of ellipses and external references, as commonly found in User-Generated Content (UGC).

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based features (namely information about the speaker and dialogue moves) was beneficial [85]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [73] for document-wise parsing and by [96] for POS tagging.

Taking the context into account requires new inference methods able to share information between sentences as well as new learning methods capable of finding out which information is to be made available, and where. Integrating contextual information at all steps of an NLP pipeline is among the main research questions addressed in this research strand. In the short term, we focus on morphological and syntactic disambiguation within close-world scenarios, as found in video games and domain-specific UGC. In the long term, we investigate the integration of linguistically motivated semantic information into joint learning models.

From a more general perspective, contexts may take many forms and require imagination to discern them, get useful data sets, and find ways to exploit them. A context may be a question associated with an answer, a rating associated with a comment (as provided by many web services), a thread of discussions (e-mails, social media, digital assistants, chatbots—on which see below–), but also meta data about some situation (such as discussions between gamers in relation with the state of the game) or multiple points of views (pictures and captions, movies and subtitles). Even if the relationship between a language production and its context is imprecise and indirect, it is still a valuable source of information, notwithstanding the need for less supervised machine learning techniques (cf. the use of LSTM neural networks by Google to automatically suggest replies to emails).

3.2.3. Information and knowledge extraction

The use of local contexts as discussed above is a new and promising approach. However, a more traditional notion of global context or world knowledge remains an open question and still raises difficult issues. Indeed, many aspects of language such as ambiguities and ellipsis can only be handled using world knowledge. Linked Open Data (LODs) such as DBpedia, WordNet, BabelNet, or Framebase provide such knowledge and we plan to exploit them.

[0]Semantic graph parsing, evaluated on biomedical data, speech and opinion. We ranked 1 in a joint effort with the Stanford NLP team

However, each specialised domain (economy, law, medicine...) exhibits its own set of concepts with associated terms. This is also true of communities (e.g. on social media), and it is even possible to find communities discussing the same topics (e.g. immigration) with very distinct vocabularies. Global LODs weakly related to language may be too general and not sufficient for a specific language variant. Following and extending previous work in ALPAGE, we put an emphasis on information acquisition from corpora, including error mining techniques in parsed corpora (to detect specific usages of a word that are missing in existing resources), terminology extraction, and word clustering.

Word clustering is of specific importance. It relies on the distributional hypothesis initially formulated by Harris, which states that words occurring in similar contexts tend to be semantically close. The latest developments of these ideas (with word2vec or GloVe) have led to the embedding of words (through vectors) in low-dimensional semantic spaces. In particular, words that are typical of several communities (see above) can be embedded in a same semantic space in order to establish mappings between them. It is also possible in such spaces to study static configurations and vector shifts with respect to variables such as time, using topological theories (such as pretopology), for instance to explore shifts in meaning over time (cf. the ANR project Profiterole concerning ancient French texts) or between communities (cf. the ANR project SoSweet). It is also worth mentioning on-going work (in computational semantics) whose goal is to combine word embeddings to embed expressions, sentences, paragraphs or even documents into semantic spaces, e.g. to explore the similarity of documents at various time periods.

Besides general knowledge about a domain, it is important to detect and keep trace of more specific pieces of information when processing a document and maintaining a context, especially about (recurring) Named Entities (persons, organisations, locations...) —something that is the focus of future work in collaboration with Patrice Lopez on named entity detection in scientific texts. Through the co-supervision of a PhD funded by the LabEx EFL (see below), we are also involved in pronominal coreference resolution (finding the referent of pronouns). Finally, we plan to continue working on deeper syntactic representations (as initiated with the Deep Sequoia Treebank), thus paving the way towards deeper semantic representations. Such information is instrumental when looking for more precise and complete information about who does what, to whom, when and where in a document. These lines of research are motivated by the need to extract useful contextual information, but it is also worth noting their strong potential in industrial applications.

3.3. Computational Modelling of Linguistic Variation

NLP and DH tools and resources are very often developed for contemporary, edited, non-specialised texts, often based on journalistic corpora. However, such corpora are not representative of the variety of existing textual data. As a result, the performance of most NLP systems decreases, sometimes dramatically, when faced with non-contemporary, non-edited or specialised texts. Despite the existence of domain-adaptation techniques and of robust tools, for instance for social media text processing, dealing with linguistic variation is still a crucial challenge for NLP and DH.

Linguistic variation is not a monolithic phenomenon. Firstly, it can result from different types of processes, such as variation over time (diachronic variation) and variation correlated with sociological variables (sociolinguistic variation, especially on social networks). Secondly, it can affect all components of language, from spelling (languages without a normative spelling, spelling errors of all kinds and origins) to morphology/syntax (especially in diachrony, in texts from specialised domains, in social media texts) and semantics/pragmatics (again in diachrony, for instance). Finally, it can constitute a property of the data to be analysed or a feature of the data to be generated (for instance when trying to simplify texts for increasing their accessibility for disabled and/or non-native readers).

Nevertheless, despite this variability in variation, the underlying mechanisms are partly comparable. This motivates our general vision that many generic techniques could be developed and adapted to handle different types of variation. In this regard, three aspects must be kept in mind: spelling variation (human errors, OCR/HTR errors, lack of spelling conventions for some languages...), lack or scarcity of parallel data aligning “variation-affected” texts and their “standard/edited” counterpart, and the sequential nature of the problem at hand. We will therefore explore, for instance, how unsupervised or weakly-supervised techniques

could be developed and feed dedicated sequence-to-sequence models. Such architectures could help develop "normalisation" tools adapted, for example, to social media texts, texts written in ancient/dialectal varieties of well-resourced languages (e.g. Old French texts), and OCR/HTR system outputs.

Nevertheless, the different types of language variation will require specific models, resources and tools. All these directions of research constitute the core of our second research strand described in this section.

3.3.1. Theoretical and empirical synchronic linguistics

Permanent members involved: all

We aim to explore computational models to deal with language variation. It is important to get more insights about language in general and about the way humans apprehend it. We will do so in at least two directions, associating computational linguistics with formal and descriptive linguistics on the one hand (especially at the morphological level) and with cognitive linguistics on the other hand (especially at the syntactic level).

Recent advances in morphology rely on quantitative and computational approaches and, sometimes, on collaboration with descriptive linguists—see for instance the special issue of the *Morphology* journal on "computational methods for descriptive and theoretical morphology", edited and introduced by [70]. In this regard, ALMAnaCH members have taken part in the design of quantitative approaches to defining and measuring morphological complexity and to assess the internal structure of morphological systems (inflection classes, predictability of inflected forms...). Such studies provide valuable insights on these prominent questions in theoretical morphology. They also improve the linguistic relevance and the development speed of NLP-oriented lexicons, as also demonstrated by ALMAnaCH members. We shall therefore pursue these investigations, and orientate them towards their use in diachronic models (see section 3.3.3).

Regarding cognitive linguistics, we have the perfect opportunity with the starting ANR-NSF project "Neuro-Computational Models of Natural Language" (NCM-NL) to go in this direction, by examining potential correlations between medical imagery applied on patients listening to a reading of "Le Petit Prince" and computation models applied on the novel. A secondary prospective benefit from the project will be information about processing evolution (by the patients) along the novel, possibly due to the use of contextual information by humans.

3.3.2. Sociolinguistic variation

Because language is central in our social interactions, it is legitimate to ask how the rise of digital content and its tight integration in our daily life has become a factor acting on language. This is even more actual as the recent rise of novel digital services opens new areas of expression, which support new linguistic behaviours. In particular, social media such as Twitter provide channels of communication through which speakers/writers use their language in ways that differ from standard written and oral forms. The result is the emergence of new language varieties.

A very similar situation exists with regard to historical texts, especially documentary texts or graffiti but even literary texts, that do not follow standardised orthography, morphology or syntax.

However, NLP tools are designed for standard forms of language and exhibit a drastic loss of accuracy when applied to social media varieties or non-standardised historical sources. To define appropriate tools, descriptions of these varieties are needed. However, to validate such descriptions, tools are also needed. We address this chicken-and-egg problem in an interdisciplinary fashion, by working both on linguistic descriptions and on the development of NLP tools. Recently, socio-demographic variables have been shown to bear a strong impact on NLP processing tools (see for instance [80] and references therein). This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria project-team Dante), we will study how these variables can be factored out by our models and, in a second step, how they can be accurately predicted from sources lacking these kinds of featured descriptions.

3.3.3. Diachronic variation

Language change is a type of variation pertaining to the diachronic axis. Yet any language change, whatever its nature (phonetic, syntactic...), results from a particular case of synchronic variation (competing phonetic realisations, competing syntactic constructions...). The articulation of diachronic and synchronic variation is influenced to a large extent by both language-internal factors (i.e. generalisation of context-specific facts) and/or external factors (determined by social class, register, domain, and other types of variation).

Very few computational models of language change have been developed. Simple deterministic finite-state-based phonetic evolution models have been used in different contexts. The PIElexicon project [90] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on an idiosyncratic and unacceptable reconstruction of the Proto-Indo-European language. Probabilistic finite-state models have also been used for automatic cognate detection and proto-form reconstruction, for example by [71] and [81]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

In ALMAnaCH, our goal is to work on modelling phonetic, morphological and lexical diachronic evolution, with an emphasis on computational etymological research and on the computational modelling of the evolution of morphological systems (morphological grammar and morphological lexicon). These efforts will be in direct interaction with sub-strand 3b (development of lexical resources). We want to go beyond the above-mentioned purely phonetic models of language and lexicon evolution, as they fail to take into account a number of crucial dimensions, among which: (1) spelling, spelling variation and the relationship between spelling and phonetics; (2) synchronic variation (geographical, genre-related, etc.); (3) morphology, especially through intra-paradigmatic and inter-paradigmatic analogical levelling phenomena, (4) lexical creation, including via affixal derivation, back-formation processes and borrowings.

We apply our models to two main tasks. The first task, as developed for example in the context of the ANR project Profiterole, consists in predicting non-attested or non-documented words at a certain date based on attestations of older or newer stages of the same word (e.g., predicting a non-documented Middle French word based on its Vulgar Latin and Old French predecessors and its Modern French successor). Morphological models and lexical diachronic evolution models will provide independent ways to perform the same predictions, thus reinforcing our hypotheses or pointing to new challenges.

The second application task is computational etymology and proto-language reconstruction. Our lexical diachronic evolution models will be paired with semantic resources (wordnets, word embeddings, and other corpus-based statistical information). This will allow us to formally validate or suggest etymological or cognate relations between lexical entries from different languages of a same language family, provided they are all inherited. Such an approach could also be adapted to include the automatic detection of borrowings from one language to another (e.g. for studying the non-inherited layers in the Ancient Greek lexicon). In the longer term, we will investigate the feasibility of the automatic (unsupervised) acquisition of phonetic change models, especially when provided with lexical data for numerous languages from the same language family.

These lines of research will rely on etymological data sets and standards for representing etymological information (see Section 3.4.2).

Diachronic evolution also applies to syntax, and in the context of the ANR project Profiterole, we are beginning to explore more or less automatic ways of detecting these evolutions and suggest modifications, relying on fine-grained syntactic descriptions (as provided by meta-grammars), unsupervised sentence clustering (generalising previous works on error mining, cf. [6]), and constraint relaxation (in meta-grammar classes). The underlying idea is that a new syntactic construction evolves from a more ancient one by small, iterative modifications, for instance by changing word order, adding or deleting functional words, etc.

3.3.4. Accessibility-related variation

Language variation does not always pertain to the textual input of NLP tools. It can also be characterised by their intended output. This is the perspective from which we investigate the issue of text simplification (for a recent survey, see for instance [109]). Text simplification is an important task for improving the accessibility to information, for instance for people suffering from disabilities and for non-native speakers learning a given

language [91]. To this end, guidelines have been developed to help writing documents that are easier to read and understand, such as the FALC ("Facile À Lire et à Comprendre") guidelines for French. [0]

Fully automated text simplification is not suitable for producing high-quality simplified texts. Besides, the involvement of disabled people in the production of simplified texts plays an important social role. Therefore, following previous works [79], [103], our goal will be to develop tools for the computer-aided simplification of textual documents, especially administrative documents. Many of the FALC guidelines can only be linguistically expressed using complex, syntactic constraints, and the amount of available "parallel" data (aligned raw and simplified documents) is limited. We will therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [68]). Lexical simplification, another aspect of text simplification [86], [92], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

Accessibility can also be related to the various presentation forms of a document. This is the context in which we have initiated the OPALINE project, funded by the *Programme d'Investissement d'Avenir - Fonds pour la Société Numérique*. The objective is for us to further develop the GROBID text-extraction suite [0] in order to be able to re-publish existing books or dictionaries, available in PDF, in a format that is accessible by visually impaired persons.

3.4. Modelling and Development of Language Resources

Language resources (raw and annotated corpora, lexical resources, etc.) are required in order to apply any machine learning technique (statistical, neural, hybrid) to an NLP problem, as well as to evaluate the output of an NLP system.

In data-driven, machine-learning-based approaches, language resources are the place where linguistic information is stored, be it implicitly (as in raw corpora) or explicitly (as in annotated corpora and in most lexical resources). Whenever linguistic information is provided explicitly, it complies to guidelines that formally define which linguistic information should be encoded, and how. Designing linguistically meaningful and computationally exploitable ways to encode linguistic information within language resources constitutes the first main scientific challenge in language resource development. It requires a strong expertise on both the linguistic issues underlying the type of resource under development (e.g. on syntax when developing a treebank) and the NLP algorithms that will make use of such information.

The other main challenge regarding language resource development is a consequence of the fact that it is a costly, often tedious task. ALMAnaCH members have a long track record of language resource development, including by hiring, training and supervising dedicated annotators. But a manual annotation can be speeded up by automatic techniques. ALMAnaCH members have also work on such techniques, and published work on approaches such as automatic lexical information extraction, annotation transfer from a language to closely related languages, and more generally on the use of pre-annotation tools for treebank development and on the impact of such tools on annotation speed and quality. These techniques are often also relevant for Research strand 1. For example, adapting parsers from one language to the other or developing parsers that work on more than one language (e.g. a non-lexicalised parser trained on the concatenation of treebanks from different languages in the same language family) can both improve parsing results on low-resource languages and speed up treebank development for such languages.

3.4.1. Construction, management and automatic annotation of Text Corpora

Corpus creation and management (including automatic annotation) is often a time-consuming and technically challenging task. In many cases, it also raises scientific issues related for instance with linguistic questions

[0] Please click here for an archived version of these guidelines (at the time this footnote is begin written, the original link does not seem to work any more).

[0] https://github.com/kermitt2/grobid

(what is the elementary unit in a text?) as well as computer-science challenges (for instance when OCR or HTR are involved). It is therefore necessary to design a work-flow that makes it possible to deal with data collections, even if they are initially available as photos, scans, wikipedia dumps, etc.

These challenges are particularly relevant when dealing with ancient languages or scripts where fonts, OCR techniques, language models may be not extant or of inferior quality, as a result, among others, of the variety of writing systems and the lack of textual data. We will therefore work on improving print OCR for some of these languages, especially by moving towards joint OCR and language models. Of course, contemporary texts can be often gathered in very large volumes, as we already do within the ANR project SoSweet, resulting in different, specific issues.

ALMAnaCH pays a specific attention to the re-usability [0] of all resources produced and maintained within its various projects and research activities. To this end, we will ensure maximum compatibility with available international standards for representing textual sources and their annotations. More precisely we will take the TEI (*Text Encoding Initiative*) guidelines as well the standards produced by ISO committee TC 37/SC 4 as essential points of reference.

From our ongoing projects in the field of Digital Humanities and emerging initiatives in this field, we observe a real need for complete but easy work-flows for exploiting corpora, starting from a set of raw documents and reaching the level where one can browse the main concepts and entities, explore their relationship, extract specific pieces of information, always with the ability to return to (fragments of) the original documents. The pieces of information extracted from the corpora also need to be represented as knowledge databases (for instance as RDF "linked data"), published and linked with other existing databases (for instance for people and locations).

The process may be seen as progressively enriching the documents with new layers of annotations produced by various NLP modules and possibly validated by users, preferably in a collaborative way. It relies on the use of clearly identified representation formats for the annotations, as advocated within ISO TC 37/SC 4 standards and the TEI guidelines, but also on the existence of well-designed collaborative interfaces for browsing, querying, visualisation, and validation. ALMAnaCH has been or is working on several of the NLP bricks needed for setting such a work-flow, and has a solid expertise in the issues related to standardisation (of documents and annotations). However, putting all these elements in a unified work-flow that is simple to deploy and configure remains to be done. In particular, work-flow and interface should maybe not be dissociated, in the sense that the work-flow should be easily piloted and configured from the interface. An option will be to identify pertinent emerging platforms in DH (such as Transkribus) and to propose collaborations to ensure that NLP modules can be easily integrated.

It should be noted that such work-flows have actually a large potential besides DH, for instance for exploiting internal documentation (for a company) or exploring existing relationships between entities.

3.4.2. Development of Lexical Resources

ALPAGE, the Inria predecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [5], [1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [101], [97], [111] and parsing, and some of the lexical resource development will be targeted towards the improvement of NLP tools. They will also play a central role for studying diachrony in the lexicon, for example for Ancient to Contemporary French in the context of the Profiterole project. They will also be one of the primary sources of linguistic information for augmenting language models used in OCR systems for ancient scripts, and will allow us to develop automatic annotation tools (e.g. POS taggers) for low-resourced languages (see already [113]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

[0] From a larger point of view we intend to comply with the so-called FAIR principles (http://force11.org/group/fairgroup/fairprinciples).

Therefore, an important effort towards the development of new morphological lexicons will be initiated, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we will try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [100], [83], [102], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [98], [104]. A new line of research will be to integrate the diachronic axis by linking lexicons that are in diachronic relation with one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty will be the integration of etymological information in these lexical resources, which requires the formalisation, the standardisation, and the extraction of etymological information from OCRised dictionaries or other electronic resources, as well as the automatic generation of candidate etymologies. These directions of research are already investigated in ALMAnaCH [83], [102].

An underlying effort for this research will be to further the development of the GROBID-dictionaries software, which provides cascading CRF (Conditional Random Fields) models for the segmentation and analysis of existing print dictionaries. The first results we have obtained have allowed us to set up specific collaborations to improve our performances in the domains of a) recent general purpose dictionaries such as the Petit Larousse (Nénufar project, funded by the DGLFLF in collaboration with the University of Montpellier), b) etymological dictionaries (in collaboration with the Berlin Brandenburg Academy of sciences) and c) patrimonial dictionaries such as the Dictionnaire Universel de Basnage (an ANR project, including a PhD thesis at ALMAnaCH, has recently started on this topic in collaboration with the University of Grenoble-Alpes and the University Sorbonne Nouvelle in Paris).

In the same way as we signalled the importance of standards for the representation of interoperable corpora and their annotations, we will keep making the best use of the existing standardisation background for the representation of our various lexical resources. There again, the TEI guidelines play a central role, and we have recently participated in the "TEI Lex 0" initiative to provide a reference subset for the "Dictionary" chapter of the guidelines. We are also responsible, as project leader, of the edition of the new part 4 of the ISO standard 24613 (LMF, Lexical Markup Framework) [94] dedicated to the definition of the TEI serialisation of the LMF model (defined in ISO 24613 part 1 'Core model', 2 'Machine Readable Dictionaries' and 3 'Etymology'). We consider that contributing to standards allows us to stabilise our knowledge and transfer our competence.

3.4.3. Development of Annotated Corpora

Along with the creation of lexical resources, ALMAnaCH is also involved in the creation of corpora either fully manually annotated (gold standard) or automatically annotated with state-of-the-art pipeline processing chains (silver standard). Annotations will either be only morphosyntactic or will cover more complex linguistic levels (constituency and/or dependency syntax, deep syntax, maybe semantics). Former members of the ALPAGE project have a renowned experience in those aspects (see for instance [107], [93], [106], [88]) and will participate to the creation of valuable resources originating from the historical domain genre.

Under the auspices of the ANR Parsiti project, led by ALMAnaCH (PI: DS), we aim to explore the interaction of extra-linguistic context and speech acts. Exploiting extra-linguistics context highlights the benefits of expanding the scope of current NLP tools beyond unit boundaries. Such information can be of spatial and temporal nature, for instance. They have been shown to improve Entity Linking over social media streams [76]. In our case, we decided to focus on a closed world scenario in order to study context and speech acts interaction. To do so, we are developing a multimodal data set made of live sessions of a first person shooter video game (Alien vs. Predator) where we transcribed all human players interactions and face expressions streamlined with a log of all in-game events linked to the video recording of the game session, as well as the recording of the human players themselves. The in-games events are ontologically organised and enable the modelling of the extra-linguistics context with different levels of granularity. Recorded over many games sessions, we already transcribed over 2 hours of speech that will serve as a basis for exploratory work, needed for the prototyping of our context-enhanced NLP tools. In the next step of this line of work, we will focus on enriching this data set with linguistic annotations, with an emphasis on co-references resolutions and predicate

argument structures. The midterm goal is to use that data set to validate a various range of approaches when facing multimodal data in a close-world environment.

Auctus Team

3. Research Program

3.1. Analysis and modelling of human behavior

3.1.1. Scientific Context

The purpose of this axis is to provide metrics to assess human behavior. We place ourselves here from the point of view of the human being and more precisely of the industrial operator. We assume the following working hypotheses: the operator's task and environmental conditions are known and circumscribed; the operator is trained in the task, production tools and safety instructions; the task is repeated with more or less frequent intervals. We focus our proposals on assessing:

- the physical and cognitive fragility of operators in order to meet assistance needs;
- cognitive biases and physical constraints leading to a loss of operator safety;
- ergonomic, performance and acceptance of the production tool.

In the industrial context, the fields that best answer these questions are work ergonomics and cognitive sciences. Scientists typically work on 4 axes: physiological/biomechanical, cognitive, psychological and sociological. More specifically, we focus on biomechanical, cognitive and psychological aspects, as described by the ANACT [24], [26]. The aim here is to translate these factors into metrics, optimality criteria or constraints in order to implement them in our methodologies for analysis, design and control of the collaborative robot.

To understand our desired contributions in robotics, we must review the current state of ergonomic workstation evaluation, particularly at the biomechanical level. The ergonomist evaluates the gesture through the observation of workstations and, generally, through questionnaires. This requires long periods of field observation, followed by analyses based on ergonomic grids (e.g. RULA [42], REBA [32], LUBA [37], OWAS [36], ROSA [60],...). Until then, the use of more complex measurement systems was reserved for laboratories, particularly biomechanical laboratories. The appearance of inexpensive sensors such as IMUs (Inertial Measurement Units) or RGB-D cameras makes it possible to consider a digitalized, and therefore objective, observation of the gesture, postures and more generally of human movement. Thanks to these sensors, which are more or less intrusive, it is now possible to permanently install observation systems on production lines. This completely changes paradigms and opens the door to longitudinal observations. It should be noted that this is comparable to the evolution of maintenance, which becomes predictive.

On the strength of this new paradigm, *ergonomic robotics* has recently taken an interest in this type of evaluation to adapt the robot's movements in order to reduce ergonomic risk scores. This approach complements the more traditional approaches that only consider the performance of the action produced by the human in interaction with the robot. However, we must go further. Indeed, the ergonomic criteria are based on the principle that the comfort positions are distant from the human articular stops. In addition, the notation must be compatible with an observation of the human being through the eye of the ergonomist. In practice, evaluations are inaccurate and subjective [63]. Moreover, they are made for quasi-static human positions without taking into account the evolution of the person's physical, physiological and psychological state. The repetition of gestures, the solicitation of muscles and joints is one of the questions that must complete these analyses. One of the methods used by ergonomists to limit biomechanical exposures is to increase variations in motor stress by rotating tasks [61]. However, this type of extrinsic method is not always possible in the industrial context [40].

One of Auctus' objectives is to show how, through a cobot, the operator's environment can be varied to encourage more appropriate motor strategies. To do so, we must focus on a field of biomechanics that studies the intrinsic variability of the motor system allowed by the joint redundancy of the human body. This motor variability refers to the natural alternation of postures, movements and muscle activity observed in the individual to respond to a requested task [61]. This natural variation leads to differences between the motor coordinates used by individuals, which evokes the notion of motor strategy [33].

As shown by the cognitive dimension of ergonomics (see above), we believe that some of these motor strategies are a physically quantifiable reflection of the operator's cognitive state. For example, fatigue [57] and its anticipation or the manual expertise (dexterous and cognitive) of the operator which allows him to anticipate his movements over long periods of time in order to preserve his body, his performance and his pain.

3.1.2. Methodology

How can we observe, understand and quantify these human motor strategies to better design and control the behavior of the cobotic assistant? When we study the systems of equations considered (kinematic, static, dynamic, musculoskeletal), several problems appear and explain our methodological choices:

- the large dimensions of the problems to be considered, due to joint, muscle and placement redundancy,
- the variabilities of the parameters, for example: physiological (consider not an operator, but a set of operators), geometric (consider a set of possible placements of the operator) and static (consider a set of forces that the operator must produce);
- the uncertainties of measurement, model approximation.

The idea is to start from a description of redundant workspaces (geometric, static, dynamic...). To do this, we use set theory approaches, based on interval analysis [3], [48], which allow us to respond to the uncertainties and variability issues previously mentioned. In addition, one of the advantages of these techniques is that they allow the results to be certified, which is essential to address safety issues. Some members of the team has already achieved success in mechanical design for performance certification and robot design [44]. The adaptation of these approaches allows us to obtain a mapping of ergonomic and efficient movements in which we can project the operators' motor strategies and thus define a metric quantifying the sensorimotor commands chosen with regard to the cognitive criteria studied.

It is therefore necessary to:

- propose new indices linking different types of performance (ergonomic biomechanical robotics, but also influence of fatigue, stress, level of expertise on the evolution of performance);
- divide the gesture into homogeneous phases: this process is complex and depends on the type of index used and the techniques used. We are exploring several ways: inverse optimal control, learning methods, or the use of techniques from signal processing.
- develop interval extensions of the identified indices. These indices are not necessarily the result of a direct model, and algorithms need to be developed or adapted (calculation of manipulability, UCM, etc.).
- Aggregate proposals into a dedicated interval analysis library (use of and contribution to the existing ALIAS-Inria and the open source IBEX library).

The originality and contribution of the methodology is to allow an analysis taking into account in the same model the measurement uncertainties (important for on-site use of analytical equipment), the variability of tasks and trajectories, and the physiological characteristics of the operators.

Other avenues of research are being explored, particularly around the inverse optimal control [49] which allows us to project human movement on the basis of performance indices and thus to offer a possible interpretation in the analysis of behaviors.

We also use automatic classification techniques: 1) to propose cognitive models that will be learned experimentally 2) for segmentation or motion recognition, for example by testing Reservoir Computing [34] approaches.

3.2. Operator / robot coupling

3.2.1. Scientific Context

Thanks to the progress made in recent years in the field of p-HRI (Physical Human-Robot Interaction), robotic systems are beginning to operate in the same workspace as humans, which is profoundly changing industrial

issues and allowing a wide variety of human-robot coupling solutions to be considered to perform the same task [25]. Different types of interactions exist. They can be classified in different ways: according to the degree of autonomy of the robot and its proximity to the user [31] with particularities for "wearable robots" [30], [29], or for collaborative robotics [62], or according to the role of the human being [58]. From a cognitive point of view, classifications are more concerned with autonomy, the complexity of information processing and the type of communication and representation of the human being by the robot [47], [64].

We proposed a classification of cobotic systems according to the configuration of the schema of interactions between humans, robots and the environment [45], [55].

The parameters of the coupling being numerous and complex, the determination of the most appropriate type of coupling for a type of problem is an open problem [50], [2], [46], [41]. The traditional approach consists in trying to identify and classify the various possible options and to select the one that seems most relevant with regard to the feasibility, efficiency, budget envelope and acceptability of the operator. One of the main objectives of our research project is to define a typology of cobots or cobotic systems in order to specify the methodology for developing the best solution: what are the criteria for defining the best robotic architecture, what type of coupling, what autonomy of the robot, what role for the operator, what risks for the human, what overall performance? These are the key issues that need to be addressed. To meet this methodological need, we propose an approach guided by experience on use cases obtained thanks to our industrial partners.

3.2.2. Methodology

Task analysis and human behavior modelling, discussed in the previous sections, should help to characterize the different types of coupling and interaction modalities, their advantages and disadvantages, in order to assist in the decision-making process. One of the ideas we would like to develop is to try to break down the task into a sequence of elementary gestures corresponding to simple motor actions performed in a clearly identified context and to evaluate for each of them the degree of feasibility in automatic mode or in robot assistance mode. The assessment must take into account a large number of parameters that relate to physical interactions, human-robot communication, reliability and human factors, including acceptability and impact on the valuation or devaluation of the operator's work. Concerning the evaluation of human factors, we have already begun to work on the subject within the more general framework of human systems interactions by operating Bayesian networks, drawing inspiration from the work of [28], [56].

The adoption of assessment criteria for a single domain (e. g. robotics or ergonomics) cannot guarantee that the performance of this coupling will be maximized. From design to evaluation, cross-effects must be constantly considered:

- impact of the cobot design on the user's performance: intuitiveness, adaptation to intra- and inter-individual variations, affordance, stress factors (noise, vibrations,...), fatigue factors (control laws, necessary attention,...) and motivation factors (effectiveness, efficiency, aesthetics,...);
- impact of user performance on cobot exploitation: risks of human error (attention error, perseveration, circumvention of procedures, syndrome outside the loop) [28].

In addition to purely physical assistance, some cobotic systems are designed to assist the operator in his decision-making. The issues of trust, acceptance, sharing of representations and co-construction of a shared awareness of the situation are then to be addressed [59].

3.3. Design of cobotic systems

3.3.1. Architectural design

Is it necessary to cobotize, robotize or assist the human being? Which mechanical architecture meets the task challenges (a serial cobot, a specific mechanism, an exoskeleton)? What type of interaction (H/R cohabitation, comanipulation, teleoperation)? These questions are the first requests from our industrial partners. For the moment, we have few comprehensive methodological answers to provide them. Choosing a collaborative robot architecture is a difficult problem [38]. It is all the more when the questions are approached from both a cognitive ergonomics and robotics perspective. There are indeed major methodological and conceptual

differences in these areas. It is therefore necessary to bridge these representational gaps and to propose an approach that takes into consideration the expectations of the robotician to model and formalize the general properties of a cobotic system as well as those of the ergonomist to define the expectations in terms of an assistance tool.

To do this, we propose a user-centered design approach, with a particular focus on human-system interactions. From a methodological point of view, this requires first of all the development of a structured experimental approach aimed at characterizing the task to be carried out through a "system" analysis but also at capturing the physical markers of its realization: movements and efforts required, ergonomic stress. This characterization must be done through the prism of the systematic study of the exchange of information (and their nature) by humans in their performance of the considered task. On the basis of these analyses, the main challenge is to define a decision support tool for the choice of the robotic architecture and for the specifications of the role assigned to the robot and the operator as well as their interactions.

The evolution of the chosen methodology is for the moment empirical, based on the user cases regularly treated in the team (see sections on contracts and partnerships).

It can be summarized for the moment as:

- identify difficult jobs on industrial sites. This is done through visits and exchanges with our partners (manager, production manager, ergonomist...);
- select some of them, then observe the human in its ecological environment. Our tools allow us to produce a motion analysis, currently based on ergonomic criteria. In parallel we carry out a physical evaluation of the task in terms of expected performance and an evaluation of the operator by means of questionnaires.
- Synthesize these first results to deduce the robotic architectures to be initiated, the key points of human-robot interaction to be developed, the difficulties in terms of human factors to be taken into account.

In addition, the different human and task analyses take advantage of the different expertise available within the team. We would like to gradually introduce the evaluation criteria presented above. Indeed, the team has already worked on the current dominant approach: the use of a virtual human to design the cobotic cell through virtual tools [1]. However, the very large dimensions of the problems treated (modelling of the body's ddl and the constraints applied to it) makes it difficult to carry out a certified analysis. We then choose to go through the calculation of the body's workspace, representing its different performances, which is not yet done in this field. The idea here is to apply set theory approaches, using interval analysis and already discussed in section 3.1.2 . The goal is then to extend to intervals the constraints played in virtual reality during the simulation. This would allow the operator to check his trajectories and scenarios not only for a single case study but also for sets of cases. For example, it can be verified that, regardless of the bounded sets of simulated operator physiologies, the physical constraints of a simulated trajectory are not violated. Thus, the assisted design tools certify cases of use as a whole. Moreover, the intersection between the human and robot workspaces provides the necessary constraints to certify the feasibility of a task. This allows us to better design a cobotic system to integrate physical constraints. In the same way, we will look for ways in which human cognitive markers can be included in this approach.

Thus, we summarize here the contributions of the other research axes, from the analysis of human behavior in its environment for an identified task, to the choice of a mechanical architecture, via an evaluation of torque and interactions. All the previous analyses provide design constraints. This methodological approach is perfectly integrated into an Appropriate Design approach used for the dimensional design of robots, again based on interval analysis. Indeed, to the desired performance of the human-robot couple in relation to a task, it is sufficient to add the constraints limiting the difficulty of the operator's gesture as described above. The challenges are then the change of scale in models that symbiotically consider the human-robot pair, the uncertain, flexible and uncontrollable nature of human behavior and the many evaluation indices needed to describe them.

3.3.2. Control design

The control laws of collaborative robots from the major robot manufacturers differ little or not at all from the existing control laws in the field of conventional industrial robotics. Security is managed a posteriori, as an exception, by a security PLC / PC. It is therefore not an intrinsic property of the controller. This quite strongly restricts the possibilities of physical interaction [0] and collaboration and leads to sub-optimal operation of the robotic system. It is difficult in this context to envision real human-robot collaboration. Collaborative operation requires, in this case, a control calculation that integrates safety and ergonomics as a priori constraints.

The control of truly collaborative robots in an industrial context is, from our point of view, underpinned by two main issues. The first is related to the macroscopic adaptation of the robot's behaviour according to the phases of the production process. The second is related to the fine adaptation of the degree and/or nature of the robot's assistance according to the ergonomic state of the operator. If this second problem is part of a historical dynamic in robotics that consists in placing safety constraints, particularly those related to the presence of a human being, at the heart of the control problem [31] [43], [35], it is not approached from the more subtle point of view of ergonomics where the objective cannot be translated only in terms of human life or death, but rather in terms of long-term respect for their physical and mental integrity. Thus, the simple and progressive adoption by a human operator of the collaborative robot intended to assist him in his gesture requires a self-adaptation in the time of the command. This self-adaptation is a fairly new subject in the literature [51], [52].

At the macroscopic level, the task plan to be performed for a given industrial operation can be represented by a finite state machine. In order not to increase the human's cognitive load by explicitly asking him to manage transitions for the robot, we propose to develop a decision algorithm to ensure discrete transitions from one task (and the associated assistance mode) to another based on an online estimate of the current state of the human-robot couple. The associated scientific challenge requires establishing a link between the robot's involvement and a given working situation. To do so, we propose an incremental approach to learning this complex relationship. The first stage of this work will consist in identifying the general and relevant control variables to conduct this learning in an efficient and reusable way, regardless of the particular method of calculating the control action. Physically realistic simulations and real word experiments will be used to feed this learning process.
In order to handle mode transitions, we propose to explore the richness of the multi-tasking control formalism under constraints [39] in order to ensure a continuous transition from one control mode to another while guaranteeing compliance with a certain number of control constraints. Some of these constraints are based on ergonomic specifications and are dependent on the state of the robot and of the human operator, which, by nature, is difficult to predict accurately. We propose to exploit the interval analysis paradigm to efficiently formulate ergonomic constraints robust to the various existing uncertainties.
Purely discrete or reactive adaptation of the control law would make no sense given the slow dynamics of certain physiological phenomena such as fatigue. Thus, we propose to formulate the control problem as a predictive problem where the impact of the control decision at a time t is anticipated at different time horizons. This requires a prediction of human movement and knowledge of the motor variability strategies it employs. This prediction is possible on the basis of the supervision at all times of the operational objectives (task in progress) in the short term. However, it requires the use of a virtual human model and possibly a dynamic simulation to quantify the impact of these potential movements in terms of performance, including ergonomics. It is impractical to use a predictive command with simulation in the loop with an advanced virtual manikin model. We therefore propose to adapt the prediction horizon and the complexity of the corresponding model in order to guarantee a reasonable computational complexity.
The planned developments require both an approach to modelling human sensorimotor behaviour, particularly in terms of accommodating fatigue via motor variability, and validating related models in an experimental framework based on observation of movement and quantification of ergonomic performance. Experimental developments must also focus on the validation of proposed control approaches in concrete contexts. To begin with, the Woobot project related to gesture assistance for carpenters (Nassim Benhahib's thesis) and

[0] In the ISO TS 15066 technical specification on collaborative robotics, human-robot physical interaction is allowed but perceived as a situation to be avoided.

a collaboration currently being set up with Safran on assistance to operators in shrink-wrapping tasks (manual knotting) in aeronautics are rich enough background elements to support the research conducted. Collaborative research projects with PSA will also soon provide a larger set of contexts in which the proposed research can be validated.

AVIZ Project-Team

3. Research Program

3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Visualization and Data Mining. Indirectly, it inherits from other established domains such as graphic design, Exploratory Data Analysis (EDA), statistics, Artificial Intelligence (AI), Human-Computer Interaction (HCI), and Psychology.

The use of graphic representation to understand abstract data is a goal Visual Analytics shares with Tukey's Exploratory Data Analysis (EDA) [67], graphic designers such as Bertin [54] and Tufte [66], and HCI researchers in the field of Information Visualization [53].

EDA is complementary to classical statistical analysis. Classical statistics starts from a *problem*, gathers *data*, designs a *model* and performs an *analysis* to reach a *conclusion* about whether the data follows the model. While EDA also starts with a problem and data, it is most useful *before* we have a model; rather, we perform visual analysis to discover what kind of model might apply to it. However, statistical validation is not always required with EDA; since often the results of visual analysis are sufficiently clear-cut that statistics are unnecessary.

Visual Analytics relies on a process similar to EDA, but expands its scope to include more sophisticated graphics and areas where considerable automated analysis is required before the visual analysis takes place. This richer data analysis has its roots in the domain of Data Mining, while the advanced graphics and interactive exploration techniques come from the scientific fields of Data Visualization and HCI, as well as the expertise of professions such as cartography and graphic designers who have long worked to create effective methods for graphically conveying information.

The books of the cartographer Bertin and the graphic designer Tufte are full of rules drawn from their experience about how the meaning of data can be best conveyed visually. Their purpose is to find effective visual representation that describe a data set but also (mainly for Bertin) to discover structure in the data by using the right mappings from abstract dimensions in the data to visual ones.

For the last 25 years, the field of Human-Computer Interaction (HCI) has also shown that interacting with visual representations of data in a tight perception-action loop improves the time and level of understanding of data sets. Information Visualization is the branch of HCI that has studied visual representations suitable to understanding and interaction methods suitable to navigating and drilling down on data. The scientific foundations of Information Visualization come from theories about perception, action and interaction.

Several theories of perception are related to information visualization such as the "Gestalt" principles, Gibson's theory of visual perception [59] and Triesman's "preattentive processing" theory [65]. We use them extensively but they only have a limited accuracy for predicting the effectiveness of novel visual representations in interactive settings.

Information Visualization emerged from HCI when researchers realized that interaction greatly enhanced the perception of visual representations.

To be effective, interaction should take place in an interactive loop faster than 100ms. For small data sets, it is not difficult to guarantee that analysis, visualization and interaction steps occur in this time, permitting smooth data analysis and navigation. For larger data sets, more computation should be performed to reduce the data size to a size that may be visualized effectively.

In 2002, we showed that the practical limit of InfoVis was on the order of 1 million items displayed on a screen [57]. Although screen technologies have improved rapidly since then, eventually we will be limited by the physiology of our vision system: about 20 millions receptor cells (rods and cones) on the retina. Another problem will be the limits of human visual attention, as suggested by our 2006 study on change blindness in large and multiple displays [55]. Therefore, visualization alone cannot let us understand very large data sets. Other techniques such as aggregation or sampling must be used to reduce the visual complexity of the data to the scale of human perception.

Abstracting data to reduce its size to what humans can understand is the goal of Data Mining research. It uses data analysis and machine learning techniques. The scientific foundations of these techniques revolve around the idea of finding a good model for the data. Unfortunately, the more sophisticated techniques for finding models are complex, and the algorithms can take a long time to run, making them unsuitable for an interactive environment. Furthermore, some models are too complex for humans to understand; so the results of data mining can be difficult or impossible to understand directly.

Unlike pure Data Mining systems, a Visual Analytics system provides analysis algorithms and processes compatible with human perception and understandable to human cognition. The analysis should provide understandable results quickly, even if they are not ideal. Instead of running to a predefined threshold, algorithms and programs should be designed to allow trading speed for quality and show the tradeoffs interactively. This is not a temporary requirement: it will be with us even when computers are much faster, because good quality algorithms are at least quadratic in time (e.g. hierarchical clustering methods). Visual Analytics systems need different algorithms for different phases of the work that can trade speed for quality in an understandable way.

Designing novel interaction and visualization techniques to explore huge data sets is an important goal and requires solving hard problems, but how can we assess whether or not our techniques and systems provide real improvements? Without this answer, we cannot know if we are heading in the right direction. This is why we have been actively involved in the design of evaluation methods for information visualization [63], [62], [60], [61], [58]. For more complex systems, other methods are required. For these we want to focus on longitudinal evaluation methods while still trying to improve controlled experiments.

3.2. Innovation

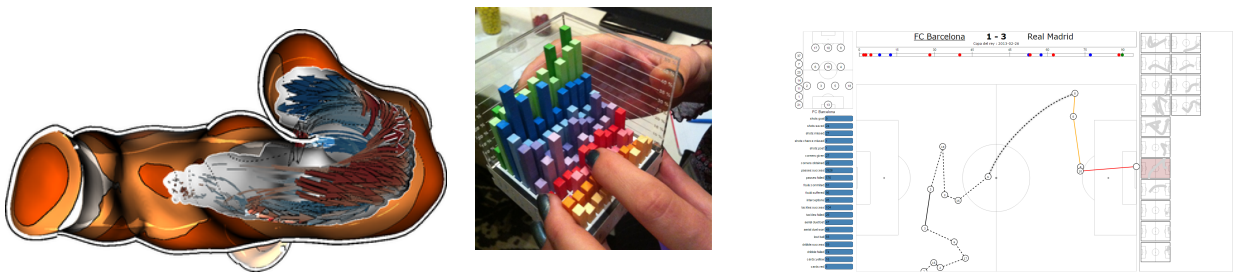


Figure 1. Example novel visualization techniques and tools developed by the team. Left: a non-photorealistic rendering technique that visualizes blood flow and vessel thickness. Middle: a physical visualization showing economic indicators for several countries, right: SoccerStories a tool for visualizing soccer games.

We design novel visualization and interaction techniques (see, for example, Figure 1). Many of these techniques are also evaluated throughout the course of their respective research projects. We cover application domains such as sports analysis, digital humanities, fluid simulations, and biology. A focus of Aviz' work is the improvement of graph visualization and interaction with graphs. We further develop individual techniques

for the design of tabular visualizations and different types of data charts. Another focus is the use of animation as a transition aid between different views of the data. We are also interested in applying techniques from illustrative visualization to visual representations and applications in information visualization as well as scientific visualization [8].

3.3. Evaluation Methods

Evaluation methods are required to assess the effectiveness and usability of visualization and analysis methods. Aviz typically uses traditional HCI evaluation methods, either quantitative (measuring speed and errors) or qualitative (understanding users tasks and activities). Moreover, Aviz is also contributing to the improvement of evaluation methods by reporting on the best practices in the field, by co-organizing workshops (BELIV 2010–2018) to exchange on novel evaluation methods, by improving our ways of reporting, interpreting and communicating statistical results, and by applying novel methodologies, for example to assess visualization literacy [3], [4].

3.4. Software Infrastructures

We want to understand the requirements that software and hardware architectures should provide to support exploratory analysis of large amounts of data. So far, "big data" has been focusing on issues related to storage management and predictive analysis: applying a well-known set of operations on large amounts of data. Visual Analytics is about exploration of data, with sometimes little knowledge of its structure or properties. Therefore, interactive exploration and analysis is needed to build knowledge and apply appropriate analyses; this knowledge and appropriateness is supported by visualizations. However, applying analytical operations on large data implies long-lasting computations, incompatible with interactions, and generates large amounts of results, impossible to visualize directly without aggregation or sampling. Visual Analytics has started to tackle these problems for specific applications but not in a general manner, leading to fragmentation of results and difficulties to reuse techniques from one application to the other. We are interested in abstracting-out the issues and finding general architectural models, patterns, and frameworks to address the Visual Analytics challenge in more generic ways.

3.5. Emerging Technologies

Figure 2. Example emerging technology solutions developed by the team for multi-display environments, wall displays, and token-based visualization.

We want to use different types of display media to empower humans to visually and interactively explore information, in order to better understand and exploit it. This includes novel display equipment and accompanying input techniques. The Aviz team specifically focuses on the exploration of the use of large displays in visualization contexts as well as emerging physical and tangible visualizations (e. g. [6], [5]). In terms of interaction modalities our work focuses on using touch and tangible interaction. Aviz participates to the Digiscope project that funds 11 wall-size displays at multiple places in the Paris area (see http://www.digiscope.fr),

connected by telepresence equipment and a Fablab for creating devices. Aviz is in charge of creating and managing the Fablab, uses it to create physical visualizations, and is also using the local wall-size display (called WILD) to explore visualization on large screens. The team also investigates the perceptual, motor and cognitive implications of using such technologies for visualization.

3.6. Psychology

More cross-fertilization is needed between psychology and information visualization. The only key difference lies in their ultimate objective: understanding the human mind vs. helping to develop better tools. We focus on understanding and using findings from psychology to inform new tools for information visualization. In many cases, our work also extends previous work in psychology. Our approach to the psychology of information visualization is largely holistic and helps bridge gaps between perception, action and cognition in the context of information visualization. Our focus includes the perception of charts in general, perception in large display environments, collaboration, perception of animations, how action can support perception and cognition, and judgment under uncertainty (e. g. [9]).

CEDAR Project-Team

3. Research Program

3.1. Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2. Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3. Multi-Model Querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4. Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example".

3.5. Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs. This research has benefitted from the arrival in the team of Mirjana Mazuran, as well as from the start of the PhD thesis of Paweł Guzewicz, co-advised by Yanlei Diao and Ioana Manolescu.

3.6. An Unified Framework for Optimizing Data Analytics

Data analytics in the cloud has become an integral part of enterprise businesses. Big data analytics systems, however, still lack the ability to take user performance goals and budgetary constraints for a task, collectively referred to as task objectives, and automatically configure an analytic job to achieve the objectives.

Our goal, is to come up with a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores as well as other runtime system parameters that best meet the task objectives. To achieve this, we also need to design a multi-objective optimizer that constructs a Pareto optimal set of job configurations for task-specific objectives, and recommends new job configurations to best meet these objectives.

CHORALE Team

3. Research Program

3.1. Task based world modeling and understanding

Executing a robotic task needs to specify a task space and a set of objective functions to be optimized. One research issue will be to define a framework allowing to represent the tasks in a generic canonic space in order to make their design and their analysis easier thanks to the control theory tools (observability, controllability, robustness...). All along the execution of the task, autonomous robotics systems have to acquire and maintain a model of the world and of the interactions between the different components involved in the task (heterogeneous robots, human beings, changes in the environment...). This model evolves in time and in space. In this research axis, we will investigate novel task-oriented world multi-layers representations (photometry, geometry, semantic) embedded in a short/long term memory framework able to handle static and dynamic events (long term mapping). A particular attention will be also paid to integrate human-robot interactions in shared environment (social skills). Another ambition of the project will be to build a bridge between model-based and machine learning methods. Understanding the world evolution is one of the key of autonomy. In this aim, we will focus on situation awareness.

3.2. Multi-sensory perception and control

Multi-sensory based perception and control is an area that starts from one single robot evolving in the environment with a set of sensors, up to a set of heterogeneous robots collaborating for the execution of a global shared task. We will address problems such as the active selection of the most suitable source of information (e.g. sensors and features) during the execution of the task and the active sensing control in order to maximize the collected information about the world modeling (including calibration and environment parameters, exogenous disturbances), allowing the task-driven sensor-based control framework to be more efficiently and robustly executed. Another issue will be the execution of a task defined by another robot or human, and to be replicated with a robot with different capabilities in perception, control and level of autonomy (i.e. heterogeneous robots). Last issues will come from the collaboration of different autonomous and heterogeneous robots in order to accomplish a shared task (mapping, robust localization, calibration, tracking, transporting, moving, ...)

CHROMA Project-Team

3. Research Program

3.1. Introduction

The Chroma team aims to deal with different issues of autonomous mobile robotics : perception, decision-making and cooperation. Figure 1 schemes the different themes and sub-themes investigated by Chroma.

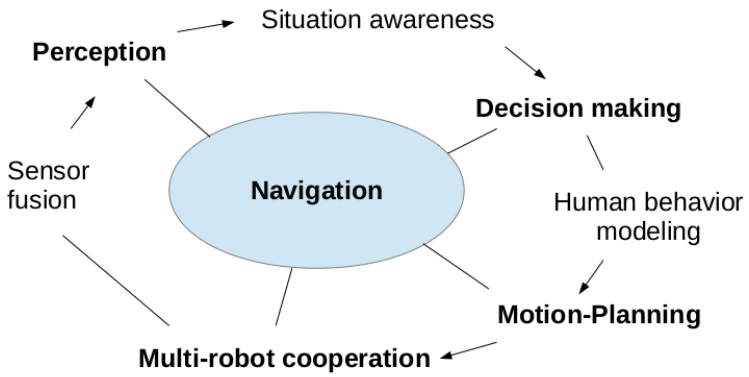


Figure 1. Research themes of the team and their relation

We present here after our approaches to address these different themes of research, and how they combine altogether to contribute to the general problem of robot navigation. Chroma pays particular attention to the problem of autonomous navigation in highly dynamic environments populated by humans and cooperation in multi-robot systems. We share this goal with other major robotic laboratories/teams in the world, such as Autonomous Systems Lab at ETH Zurich, Robotic Embedded Systems Laboratory at USC, KIT [0] (Prof Christoph Stiller lab and Prof Ruediger Dillmann lab), UC Berkeley, Vislab Parma (Prof. Alberto Broggi), and iCeiRA [0] laboratory in Taipei, to cite a few. Chroma collaborates at various levels (visits, postdocs, research projects, common publications, etc.) with most of these laboratories, see Section 9.3 .

3.2. Perception and Situation Awareness

Robust perception in open and dynamic environments populated by human beings is an open and challenging scientific problem. Traditional perception techniques do not provide an adequate solution for these problems, mainly because such environments are uncontrolled [0] and exhibit strong constraints to be satisfied (in particular high dynamicity and uncertainty). This means that **the proposed solutions have to simultaneously take into account characteristics such as real time processing, temporary occultations, dynamic changes or motion predictions**.

[0]Karlsruhe Institut fur Technologie
[0]International Center of Excellence in Intelligent Robotics and Automation Research.
[0]partially unknown and open

3.2.1. Bayesian perception

Context. Perception is known to be one of the main bottlenecks for robot motion autonomy, in particular when navigating in open and dynamic environments is subject to strong real-time and uncertainty constraints. In order to overcome this difficulty, we have proposed in the scope of the former e-Motion team, a new paradigm in robotics called "Bayesian Perception". The foundation of this approach relies on the concept of "Bayesian Occupancy Filter (BOF)" initially proposed in the Ph.D. thesis of Christophe Coue [65] and further developed in the team [0]. The basic idea is to combine a Bayesian filter with a probabilistic grid representation of both the space and the motions. It allows the filtering and the fusion of heterogeneous and uncertain sensors data, by taking into account the history of the sensors measurements, a probabilistic model of the sensors and of the uncertainty, and a dynamic model of the observed objects motions.

In the scope of the Chroma team and of several academic and industrial projects (in particular the IRT Security for autonomous vehicle and Toyota projects), we went on with the development and the extension under strong embedded implementation constraints, of our Bayesian Perception concept. This work has already led to the development of more powerful models and more efficient implementations, e.g. the *CMCDOT* (Conditional Monte Carlo Dense Occupancy Tracker) framework [89] which is still under development.

This work is currently mainly performed in the scope of the "Security for Autonomous Vehicle (SAV)" project (IRT Nanoelec), and more recently in cooperation with some Industrial Companies (see section New Results for more details on the non confidential industrial cooperation projects).

Objectives. We aim at defining a complete framework extending the Bayesian Perception paradigm to the object level. The main objective is to be simultaneously more robust, more efficient for embedded implementations, and more informative for the subsequent scene interpretation step (Figure 2 .a illustrates). Another objective is to improve the efficiency of the approach (by exploiting the highly parallel characteristic of our approach), while drastically reducing important factors such as the required memory size, the size of the hardware component, its price and the required energy consumption. This work is absolutely necessary for studying embedded solutions for the future generation of mobile robots and autonomous vehicles. We also aim at developing strong partnerships with non-academic partners in order to adapt and move the technology closer to the market.

Figure 2. a. Illustration of the Bayesian Perception Paradigm: Filtered occupancy grids, enhanced with motion estimations (vectors) and object detection (colored boxes) b. Autonomous Zoe car of Inria/Chroma.

[0]The Bayesian programming formalism developed in e-Motion, pioneered (together with the contemporary work of Thrun, Burgard and Fox [96]) a systematic effort to formalize robotics problems under Probability theory–an approach that is now pervasive in Robotics.

3.2.2. System validation

Context. Testing and validating Cyber Physical Systems which are designed for operating in various real world conditions, is both an open scientific question and a necessity for a future deployment of such systems. In particular, this is the case for Embedded Perception and Decision-making Systems which are designed for future ADAS [0] and Autonomous Vehicles. Indeed, it is unrealistic to try to be exhaustive by making a huge number of experiments in various real situations. Moreover, such experiments might be dangerous, highly time consuming, and expensive. This is why we have decided to develop appropriate *realistic simulation and statistical analysis tools* for being able to perform a huge number of tests based on some previously recorded real data and on random changes of some selected parameters (the "co-simulation" concept). Such an approach might also be used in a training step of a machine learning process. This is why simulation-based validation is getting more and more popular in automotive industry and research.

This work is performed in the scope of both the SAV [0] project (IRT Nanoelec) and of the EU Enable-S3 project; it is also performed in cooperation with the Inria team Tamis in Rennes, with the objective to integrate the Tamis "Statistical Model Checking" (SMC) approach into our validation process. We are also starting to work on this topic with the Inria team Convecs, with the objective to also integrate formal methods into our validation process.

Objectives. We started to work on this new research topic in 2017. The first objective is to build a "simulated navigation framework" for: (1) constructing realistic testing environments (including the possibility of using real experiments records), (2) developing for each vehicle a simulation model including various physical and dynamic characteristics (e.g. physics, sensors and motion control), and (3) evaluating the performances of a simulation run using appropriate statistical software tools.

The second objective is to develop models and tools for automating the Simulation & Validation process, by using a selection of relevant randomized parameters for generating large database of tests and statistical results. Then, a metric based on the use of some carefully selected "Key Performance Indicator" (KPI) has to be defined for performing a statistical evaluation of the results (e.g. by using the above-mentioned SMC approach).

3.2.3. Situation Awareness and Prediction

Context. Predicting the evolution of the perceived moving agents in a dynamic and uncertain environment is mandatory for being able to safely navigate in such an environment. We have recently shown that an interesting property of the Bayesian Perception approach is to generate short-term conservative [0] predictions on the likely future evolution of the observed scene, even if the sensing information is temporary incomplete or not available [84]. But in human populated environments, estimating more abstract properties (e.g. object classes, affordances, agent's intentions) is also crucial to understand the future evolution of the scene. This work is carried out in the scope of the Security of Autonomous Vehicle (SAV) project (IRT Nanoelec) and of several cooperative and PhD projects with Toyota and with Renault.

Objectives. The first objective is to develop an integrated approach for "Situation Awareness & Risk Assessment" in complex dynamic scenes involving multiples moving agents (e.g vehicles, cyclists, pedestrians ...), whose behaviors are most of the time unknown but predictable. Our approach relies on combining machine learning to build a model of the agent behaviors and generic motion prediction techniques (e.g. Kalman-based, GHMM , or Gaussian Processes). In the perspective of a long-term prediction we will consider the semantic level [0] combined with planning techniques.

[0] Advance Driving Assistance System
[0] Security for Autonomous Vehicles
[0] i.e. when motion parameters are supposed to be stable during a small amount of time
[0] knowledge about agent's activities and tasks

The second objective is to build a general framework for perception and decision-making in multi-robot/vehicle environments. The navigation will be performed under both dynamic and uncertainty constraints, with contextual information and a continuous analysis of the evolution of the probabilistic collision risk. Interesting published and patented results [76] have already been obtained in cooperation with Renault and UC Berkeley, by using the "Intention / Expectation" paradigm and Dynamic Bayesian Networks. We are currently working on the generalization of this approach, in order to take into account the dynamics of the vehicles and multiple traffic participants. The objective is to design a new framework, allowing us to overcome the shortcomings of rules-based reasoning approaches which often show good results in low complexity situations, but which lead to a lack of scalability and of long terms predictions capabilities.

3.2.4. Robust state estimation (Sensor fusion)

Context. In order to safely and autonomously navigate in an unknown environment, a mobile robot is required to estimate in real time several physical quantities (e.g., position, orientation, speed). These physical quantities are often included in a common state vector and their simultaneous estimation is usually achieved by fusing the information coming from several sensors (e.g., camera, laser range finder, inertial sensors). The problem of fusing the information coming from different sensors is known as the *Sensor Fusion* problem and it is a fundamental problem which plays a major role in robotics.

Objective. A fundamental issue to be investigated in any sensor fusion problem is to understand whether the state is observable or not. Roughly speaking, we need to understand if the information contained in the measurements provided by all the sensors allows us to carry out the estimation of the state. If the state is not observable, we need to detect a new observable state. This is a fundamental step in order to properly define the state to be estimated. To achieve this goal, we apply standard analytic tools developed in control theory together with some new theoretical concepts we introduced in [78] (concept of continuous symmetry). Additionally, we want to account the presence of disturbances in the observability analysis.

Our approach is to introduce general analytic tools able to derive the observability properties in the nonlinear case when some of the system inputs are unknown (and act as disturbances). We recently obtained a simple analytic tool able to account the presence of unknown inputs [80], which extends a heuristic solution derived by the team of Prof. Antonio Bicchi [60] with whom we collaborate (Centro Piaggio at the University of Pisa).

Fusing visual and inertial data. A special attention is devoted to the fusion of inertial and monocular vision sensors (which have strong application for instance in UAV navigation). The problem of fusing visual and inertial data has been extensively investigated in the past. However, most of the proposed methods require a state initialization. Because of the system nonlinearities, lack of precise initialization can irreparably damage the entire estimation process. In literature, this initialization is often guessed or assumed to be known [70]. Recently, this sensor fusion problem has been successfully addressed by enforcing observability constraints [74] and by using optimization-based approaches [77]. These optimization methods outperform filter-based algorithms in terms of accuracy due to their capability of relinearizing past states. On the other hand, the optimization process can be affected by the presence of local minima. We are therefore interested in a deterministic solution that analytically expresses the state in terms of the measurements provided by the sensors during a short time-interval.

For some years we explore deterministic solutions as presented in [79] and [81]. Our objective is to improve the approach by taking into account the biases that affect low-cost inertial sensors (both gyroscopes and accelerometers) and to exploit the power of this solution for real applications. This work is currently supported by the ANR project VIMAD [0] and experimented with a quadrotor UAV. We have a collaboration with Prof. Stergios Roumeliotis (the leader of the MARS lab at the University of Minnesota) and with Prof. Anastasios Mourikis from the University of California Riverside. Regarding the usage of our solution for real applications we have a collaboration with Prof. Davide Scaramuzza (the leader of the Robotics and Perception group at the University of Zurich) and with Prof. Roland Siegwart from the ETHZ.

[0] Navigation autonome des drones aériens avec la fusion des données visuelles et inertielles, lead by A. Martinelli, Chroma.

3.3. Navigation and cooperation in dynamic environments

In his reference book *Planning algorithms*[0] S. LaValle discusses the different dimensions that made the motion-planning problem complex, which are the number of robots, the obstacle region, the uncertainty of perception and action, and the allowable velocities. In particular, it is emphasized that complete algorithms require at least exponential time to deal with multiple robot planning in complex environments, preventing them to be scalable in practice. Moreover, dynamic and uncertain environments, as human-populated ones, expand this complexity.

In this context, we aim at **scale up decision-making in human-populated environments and in multi-robot systems, while dealing with the intrinsic limits of the robots (computation capacity, limited communication)**.

3.3.1. Motion-planning in human-populated environment

Context. Motion planning in dynamic and human-populated environments is a current challenge of robotics. Many research teams work on this topic. We can cite the Institut of robotic in Barcelone [69], the MIT [57], the Autonomous Intelligent Systems lab in Freiburg [61], or the LAAS [85]. In Chroma, we explore different issues : **integrating the risk (uncertainty) in planning processes, modeling and taking into account human behaviors and flows.**

Objective We aim to give the robot some socially compliant behaviors by anticipating the near future (trajectories of mobile obstacle in the robot's surroundings) and by integrating knowledge from psychology, sociology and urban planning. In this context, we will focus on the following 3 topics.

Risk-based planning. Unlike static or controlled environments [0] where global path planning approaches are suitable, dealing with highly dynamic and uncertain environments requires to integrate the notion of risk (risk of collision, risk of disturbance). Then, we examine how motion planning approaches can integrate this risk in the generation and selection of the paths. An algorithm called RiskRRT was proposed in the previous eMotion team. This algorithm plans goal oriented trajectories that minimize the risk estimated at each instant. It fits environments that are highly dynamic and adapts to a representation of uncertainty [93] (see Figure 3 .a for illustration). Now, we extend this principle to be adapted to various risk evaluation methods (proposed in 3.2) and various situation (highways, urban environments, even in dense traffic).

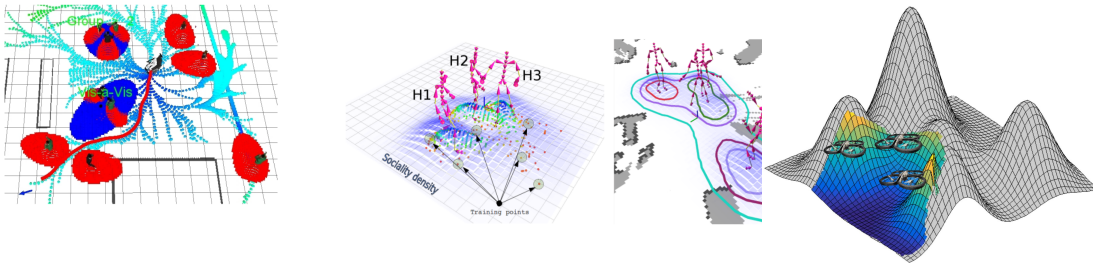


Figure 3. Illustrations of a. the Risk-RRT planning b. The human interaction space model c. Multi-UAV 3D coverage and exploration.

Recently we investigated the automatic learning of robot navigation in complex environments based on specific tasks and from visual input. We address this problem by combining computer vision, machine learning (deep-learning), and robotics path planning (see 7.5.1).

[0] Steven M. LaValle, Planning Algorithms, Cambridge University Press, 2006.
[0] known environment without uncertainty

Sharing the physical space with humans. Robots are expected to share their physical space with humans. Hence, robots need to take into account the presence of humans and to behave in a socially acceptable way. Their trajectories must be safe but also predictable, that is why they must follow social conventions, respecting proximity constraints, avoiding people interacting or joining a group engaged in conversation without disturbing. For this purpose, we proposed earlier to integrate some knowledge from the psychology domain (i.e. proxemics theory), see figure 3 .b. We aim now to integrate semantic knowledge [0] and psycho-social theories of human behavior [00] in the navigation framework we have developed for a few years (i.e. the Risk-based navigation algorithms [71], [93], [100]). These concepts were tested on our automated wheelchair (see figure 3 .c) but they have and will be adapted to autonomous cars, telepresence robots and companion robots. This work is currently supported by the ANR Valet and the ANR Hianic.

3.3.2. Decision Making in Multi-robot systems

Context. A central challenge in Chroma is to define **decision-making algorithms that scale up to large multi-robot systems**. This work takes place in the general framework of Multi-Agent Systems (MAS). The objective is to compute/define agent behaviors that provide cooperation and adaptation abilities. Solutions must also take into account the agent/robot computational limits.

We can abstract the challenge in three objectives :
i) mastering the complexity of large fleet of robots/vehicles (scalability),
ii) dealing with limited computational/memory capacity,
iii) building adaptive solutions (robustness).

Combining Decision-theoretic models and Swarm intelligence.

Over the past few years, our attempts to address multi-robot decision-making are mainly due to Multi-Agent Sequential Decision Making (MA-SDM) and Swarm Intelligence (SI). MA-SDM builds upon well-known decision-theoretic models (e.g., Markov decision processes and games) and related algorithms, that come with strong theoretical guarantees. In contrast, the expressiveness of MA-SDM models has limited scalability in face of realistic multi-robot systems [0], resulting in computational overload. On their side, SI methods, which rely on local rules – generally bio-inspired – and relating to Self-Organized Systems [0], can scale up to multiple robots and provide robustness to disturbances, but with poor theoretical guarantees [0]. Swarm models can also answer to the need of designing tractable solutions [92], but they remain not geared to express complex realistic tasks or to handle (point-to-point) communication between robots. This motivates our work to go beyond these two approaches and to combine them.

First, we plan to investigate **incremental expansion mechanisms in anytime decision-theoretic planning**, starting from local rules (from SI) to complex strategies with performance guarantees (from MA-SDM) [67]. This methodology is grounded into our research on anytime algorithms, that are guaranteed to stop at anytime while still providing a reliable solution to the original problem. It further relies on decision theoretical models and tools including: Decentralized and Partially Observable Markov Decision Processes and Games, Dynamic Programming, Distributed Reinforcement Learning and Statistical Machine Learning.

[0]B. Kuipers, The Spatial Semantic Hierarchy, Artificial Intelligence, Volume 119, Issues 1–2, May 2000, Pages 191-233
[0]Gibson, J. (1977). The theory of affordances, in Perceiving, Acting, and Knowing. Towards an Ecological Psychology. Number eds Shaw R., Bransford J. Hoboken,NJ: John Wiley & Sons Inc.
[0]Hall, E. (1966). The hidden dimension. Doubleday Anchor Books.
[0]Martin L. Puterman, Markov Decision Processes; Stuart Russell and Peter Norvig, Artificial Intelligence - A Modern Approach
[0]D. Floreano and C. Mattiussi, Bio-Inspired Artificial Intelligence - Theories, Methods, and Technologies, MIT Press, 2008.
[0]S. A. Brueckner, G. Di Marzo Serugendo, A. Karageorgos, R. Nagpal (2005). Engineering Self-Organising Systems, Methodologies and Applications. LNAI 3464 State-of-the-Art Survey, Springer book.

Second, we plan to extend the SI approach by considering **the integration of optimization techniques at the local level**. The purpose is to force the system to explore solutions around the current stabilized state – potentially a local optimum – of the system. We aim at keeping scalability and self-organization properties by not compromising the decentralized nature of such systems. Introducing optimization in this way requires to measure locally the performances, which is generally possible from local perception of robots (or using learning techniques). The main optimization methods we will consider are Local Search (Gradient Descent), Distributed Stochastic Algorithm and Reinforcement Learning. We have shown in [97] the interest of such an approach for driverless vehicle traffic optimization.

Both approaches must lead to **master the complexity** inherent to large and open multi-robot systems. Such systems are prone to combinatorial problems, in term of state space and communication, when the number of robots grows. To cope with this complexity we explore several approaches :

- Combining MA-SDM, machine learning and OR [0] techniques to deal with global-local optimization in multi-agent/robot systems. In 2016, we started a collaboration with the VOLVO Group, in Lyon, to deal with VRP problems and optimization of goods distribution using a fleet of autonomous vehicles. We also explore such a methodology in the framework of the collaboration with the team of Prof. G. Czibula (Cluj Univiersity, Romania).
- Defining heuristics by decentralizing global exact solutions. For instance we explore online stochastic-optimization planning to deal with multi-robot coverage/exploration of 3D environments, see Fig 3 .c and [42].

Beyond this methodological work, we aim to evaluate our models on benchmarks from the literature, by using simulation tools as a complement of robotic experiments. This will lead us to develop simulators, allowing to deploy tens to thousands robots in constrained environments.

Towards adaptive connected robots.

Mobile robots and autonomous vehicles are becoming more connected to one another and to other devices in the environment (concept of cloud of robots [0] and V2V/V2I connectivity in transportation systems). Such robotic systems are open systems as the number of connected entities is varying dynamically. Network of robots brought with them new problems, as the need of (online) adaption to changes in the system and to the variability of the communication.

In Chroma, we address the problem of adaptation by considering machine learning techniques and local mechanisms as discussed above (SI models). More specifically we investigate the problem of maintaining the connectivity between robots which perform dynamic version of tasks such as patrolling, exploration or transportation, i.e. where the setting of the problem is continuously changing and growing (see [86]).

In Lyon, the CITI Laboratory conducts research in many aspects of telecommunication, from signal theory to distributed computation. In this context, Chroma develops cooperations with the Inria team Agora [86] (wireless communication protocols) and with Dynamid team [63] (middlleware and cloud aspects), that we wish to reinforce in the next years.

[0] Operations Research

[0] see for instance the first International Workshop on Cloud and Robotics, 2016.

COML Team

3. Research Program

3.1. Background

In recent years, Artificial Intelligence (AI) has achieved important landmarks in matching or surpassing human level performance on a number of high level tasks (playing chess and go, driving cars, categorizing picture, etc., [31], [34], [39], [30], [36]). These strong advances were obtained by deploying on large amounts of data, massively parallel learning architectures with simple brain-inspired 'neuronal' elements. However, humans brains still outperform machines in several key areas (language, social interactions, common sense reasoning, motor skills), and are more flexible : Whereas machines require extensive expert knowledge and massive training for each particular application, humans learn autonomously over several time scales: over the developmental scale (months), humans infants acquire cognitive skills with noisy data and little or no expert feedback (weakly/unsupervised learning)[1]; over the short time scale (minutes, seconds), humans combine previously acquired skills to solve new tasks and apply rules systematically to draw inferences on the basis of extremely scarce data (learning to learn, domain adaptation, one- or zero-shot learning) [33].

The general aim of CoML, following the roadmap described in [1], is to bridge the gap in cognitive flexibility between humans and machines learning in language processing and common sense reasoning by reverse engineering how young children between 1 and 4 years of age learn from their environment. We conduct work along two axes: the first one, which we called *Developmental AI* is focused on building infant inspired machine learning algorithms. The second axis is devoted to using the developed algorithms to conduct *quantitative studies* of how infant learn across diverse environments.

3.2. Weakly/Unsupervised Learning

Much of standard machine learning is construed as regression or classification problems (mapping input data to expert-provided labels). Human infants rarely learn in this fashion, at least before going to school: they learn language, social cognition, and common sense autonomously (without expert labels) and when adults provide feedback, it is ambiguous and noisy and cannot be taken as a gold standard. Modeling or mimicking such achievement requires deploying unsupervised or weakly supervised algorithms which are less well known than their supervised counterparts.

We take inspiration from infant's landmarks during their first years of life: they are able to learn acoustic models, a lexicon, and susbtantive elements of language models and world models from raw sensory inputs. Building on previous work [3], [7], [11], we use DNN and Bayesian architectures to model the emergence of linguistic representations without supervision. Our focus is to establish how the labels in supervised settings can be replaced by weaker signals coming either from multi-modal input or from hierarchically organised linguistic levels.

At the level of phonetic representations, we study how cross-modal information (lips and self feedback from articulation) can supplement top-down lexical information in a weakly supervised setting. We use Siamese architectures or Deep CCA algorithms to combine the different views. We study how an attentional framework and uncertainty estimation can flexibly combine these informations in order to adapt to situations where one view is selectively degraded.

At the level of lexical representations, we study how audio/visual parallel information (ie. descriptions of images or activities) can help in segmenting and clustering word forms, and vice versa, help in deriving useful visual features. To achieve this, we will use architectures deployed in image captioning or sequence to sequence translation [37].

At the level of semantic and conceptual representations, we study how it is possible to learn elements of the laws of physics through the observation of videos (object permanence, solidity, spatio-temporal continuity, inertia, etc.), and how objects and relations between objects are mapped onto language.

3.3. Evaluating Machine Intelligence

Increasingly, complicated machine learning systems are being incorporated into real-life applications (e.g. self-driving cars, personal assistants), even though they cannot be formally verified, guaranteed statistically, nor even explained. In these cases, a well defined *empirical approach* to evaluation can offer interesting insights into the functioning and offer some control over these algorithms.

Several approaches exist to evaluate the 'cognitive' abilities of machines, from the subjective comparison of human and machine performance [38] to application-specific metrics (e.g., in speech, word error rate). A recent idea consist in evaluating an AI system in terms of it's *abilities* [32] , i.e., functional components within a more global cognitive architecture [35]. Psychophysical testing can offer batteries of tests using simple tasks that are easy to understand by humans or animals (e.g, judging whether two stimuli are same or different, or judging whether one stimulus is 'typical') which can be made selective to a specific component and to rare but difficult or adversarial cases. Evaluations of learning rate, domain adaptation and transfer learning are simple applications of these measures. Psychophysically inspired tests have been proposed for unsupervised speech and language learning [10], [6].

3.4. Documenting human learning

Infants learn their first language in a spontaneous fashion, across a lot of variation in amount of speech and the nature of the infant/adult interaction. In some linguistic communities, adults barely address infants until they can themselves speak. Despite these large variations in quantity and content, language learning proceeds at similar paces. Documenting such resilience is an essential step in understanding the nature of the learning algorithms used by human infants. Hence, we propose to collect and/or analyse large datasets of inputs to infants and correlate this with outcome measure (phonetic learning, vocabulary growth, syntactic learning, etc.).

DEFROST Project-Team

3. Research Program

3.1. Introduction

Our research crosses different disciplines: numerical mechanics, control design, robotics, optimisation methods and clinical applications. Our organisation aims at facilitating the team work and cross-fertilisation of research results in the group. We have three objectives (1, 2 and 3) that correspond to the main scientific challenges. In addition, we have two transverse objectives that are also highly challenging: the development of a high performance software support for the project (objective 4) and the validation tools and protocols for the models and methods (objective 5).

3.2. Objective 1: Accurate model of soft robot deformation computed in finite time

The objective is to find concrete numerical solutions to the challenge of modelling soft robots with strong real-time constraints. To solve continuum mechanics equations, we will start our research with real-time FEM or equivalent methods that were developed for soft-tissue simulation. We will extend the functionalities to account for the needs of a soft-robotic system:

- Coupling with other physical phenomenons that govern the activity of sensors and actuators (hydraulic, pneumatic, electro-active polymers, shape-memory alloys...).
- Fulfilling the new computational time constraints (harder than surgical simulation for training) and find better tradeoff between cost and precision of numerical solvers using reduced-order modelling techniques with error control.
- Exploring interactive and semi-automatic optimisation methods for design based on obtained solution for fast computation on soft robot models.

3.3. Objective 2: Model based control of soft robot behavior

The focus of this objective is on obtaining a generic methodology for soft robot feedback control. Several steps are needed to design a model based control from FEM approach:

- The fundamental question of the kinematic link between actuators, sensors, effectors and contacts using the most reduced mathematical space must be carefully addressed. We need to find efficient algorithms for real-time projection of non-linear FEM models in order to pose the control problem using the only relevant parameters of the motion control.
- Intuitive remote control is obtained when the user directly controls the effector motion. To add this functionality, we need to obtain real-time inverse models of the soft robots by optimisation. Several criteria will be combined in this optimisation: effector motion control, structural stiffness of the robot, reduce intensity of the contact with the environment...
- Investigating closed-loop approaches using sensor feedback: as sensors cannot monitor all points of the deformable structure, the information provided will only be partial. We will need additional algorithms based on the FEM model to obtain the best possible treatment of the information. The final objective of these models and algorithms is to have robust and efficient feedback control strategies for soft robots. One of the main challenge here is to ensure / prove stability in closed-loop.

3.4. Objective 3: Modeling the interaction with a complex environment

Even if the inherent mechanical compliance of soft robots makes them safer, more robust and particularly adapted to interaction with fragile environments, the contact forces need to be controlled by:

- Setting up real-time modelling and the control methods needed to pilot the forces that the robot imposes on its environment and to control the robot deformations imposed by its environment. Note that if an operative task requires to apply forces on the surrounding structures, the robot must be anchored to other structures or structurally rigidified.
- Providing mechanics models of the environment that include the uncertainties on the geometry and on the mechanical properties, and are capable of being readjusted in real-time.
- Using the visual feedback of the robot behavior to adapt dynamically the models. The observation provided in the image coupled with an inverse accurate model of the robot could transform the soft robot into sensor: as the robot deforms with the contact of the surroundings, we could retrieve some missing parameters of the environment by a smart monitoring of the robot deformations.

3.5. Objective 4: Soft Robotics Software

Expected research results of this project are numerical methods and algorithms that require high-performance computing and suitability with robotic applications. There is no existing software support for such development. We propose to develop our own software, in a suite split into three applications:

- The first one will facilitate the design of deformable robots by an easy passage from CAD software (for the design of the robot) to the FEM based simulation.
- The second one is an anticipative clinical simulator. The aim is to co-design the robotic assistance with the physicians, thanks to a realistic simulation of the procedure or the robotic assistance. This will facilitate the work of reflection on new clinical approaches prior any manufacturing.
- The third one is the control design software. It will provide the real-time solutions for soft robot control developed in the project.

3.6. Objective 5: Validation and application demonstrations

The implementation of experimental validation is a key challenge for the project. On one side, we need to validate the model and control algorithms using concrete test case example in order to improve the modelling and to demonstrate the concrete feasibility of our methods. On the other side, concrete applications will also feed the reflexions on the objectives of the scientific program.

We will build our own experimental soft robots for the validation of objectives 2 and 3 when there is no existing “turn-key” solution. Designing and making our own soft robots, even if only for validation, will help the setting-up of adequate models.

For the validation of objective 4, we will develop “anatomical soft robot”: soft robot with the shape of organs, equipped with sensors (to measure the contact forces) and actuators (to be able to stiffen the walls and recreate natural motion of soft-tissues). We will progressively increase the level of realism of this novel validation set-up to come closer to the anatomical properties.

EX-SITU Project-Team

3. Research Program

3.1. Research Program

We characterize Extreme Situated Interaction as follows:

Extreme users. We study extreme users who make extreme demands on current technology. We know that human beings take advantage of the laws of physics to find creative new uses for physical objects. However, this level of adaptability is severely limited when manipulating digital objects. Even so, we find that creative professionals—artists, designers and scientists—often adapt interactive technology in novel and unexpected ways and find creative solutions. By studying these users, we hope to not only address the specific problems they face, but also to identify the underlying principles that will help us to reinvent virtual tools. We seek to shift the paradigm of interactive software, to establish the laws of interaction that significantly empower users and allow them to control their digital environment.

Extreme situations. We develop extreme environments that push the limits of today's technology. We take as given that future developments will solve "practical" problems such as cost, reliability and performance and concentrate our efforts on interaction in and with such environments. This has been a successful strategy in the past: Personal computers only became prevalent after the invention of the desktop graphical user interface. Smartphones and tablets only became commercially successful after Apple cracked the problem of a usable touch-based interface for the iPhone and the iPad. Although wearable technologies, such as watches and glasses, are finally beginning to take off, we do not believe that they will create the major disruptions already caused by personal computers, smartphones and tablets. Instead, we believe that future disruptive technologies will include fully interactive paper and large interactive displays.

Our extensive experience with the Digiscope WILD and WILDER platforms places us in a unique position to understand the principles of distributed interaction that extreme environments call for. We expect to integrate, at a fundamental level, the collaborative capabilities that such environments afford. Indeed almost all of our activities in both the digital and the physical world take place within a complex web of human relationships. Current systems only support, at best, passive sharing of information, e.g., through the distribution of independent copies. Our goal is to support active collaboration, in which multiple users are actively engaged in the lifecycle of digital artifacts.

Extreme design. We explore novel approaches to the design of interactive systems, with particular emphasis on extreme users in extreme environments. Our goal is to empower creative professionals, allowing them to act as both designers and developers throughout the design process. Extreme design affects every stage, from requirements definition, to early prototyping and design exploration, to implementation, to adaptation and appropriation by end users. We hope to push the limits of participatory design to actively support creativity at all stages of the design lifecycle. Extreme design does not stop with purely digital artifacts. The advent of digital fabrication tools and FabLabs has significantly lowered the cost of making physical objects interactive. Creative professionals now create hybrid interactive objects that can be tuned to the user's needs. Integrating the design of physical objects into the software design process raises new challenges, with new methods and skills to support this form of extreme prototyping.

Our overall approach is to identify a small number of specific projects, organized around four themes: *Creativity, Augmentation, Collaboration* and *Infrastructure*. Specific projects may address multiple themes, and different members of the group work together to advance these different topics.

FLOWERS Project-Team

3. Research Program

3.1. Research Program

Research in artificial intelligence, machine learning and pattern recognition has produced a tremendous amount of results and concepts in the last decades. A blooming number of learning paradigms - supervised, unsupervised, reinforcement, active, associative, symbolic, connectionist, situated, hybrid, distributed learning... - nourished the elaboration of highly sophisticated algorithms for tasks such as visual object recognition, speech recognition, robot walking, grasping or navigation, the prediction of stock prices, the evaluation of risk for insurances, adaptive data routing on the internet, etc... Yet, we are still very far from being able to build machines capable of adapting to the physical and social environment with the flexibility, robustness, and versatility of a one-year-old human child.

Indeed, one striking characteristic of human children is the nearly open-ended diversity of the skills they learn. They not only can improve existing skills, but also continuously learn new ones. If evolution certainly provided them with specific pre-wiring for certain activities such as feeding or visual object tracking, evidence shows that there are also numerous skills that they learn smoothly but could not be "anticipated" by biological evolution, for example learning to drive a tricycle, using an electronic piano toy or using a video game joystick. On the contrary, existing learning machines, and robots in particular, are typically only able to learn a single pre-specified task or a single kind of skill. Once this task is learnt, for example walking with two legs, learning is over. If one wants the robot to learn a second task, for example grasping objects in its visual field, then an engineer needs to re-program manually its learning structures: traditional approaches to task-specific machine/robot learning typically include engineer choices of the relevant sensorimotor channels, specific design of the reward function, choices about when learning begins and ends, and what learning algorithms and associated parameters shall be optimized.

As can be seen, this requires a lot of important choices from the engineer, and one could hardly use the term "autonomous" learning. On the contrary, human children do not learn following anything looking like that process, at least during their very first years. Babies develop and explore the world by themselves, focusing their interest on various activities driven both by internal motives and social guidance from adults who only have a folk understanding of their brains. Adults provide learning opportunities and scaffolding, but eventually young babies always decide for themselves what activity to practice or not. Specific tasks are rarely imposed to them. Yet, they steadily discover and learn how to use their body as well as its relationships with the physical and social environment. Also, the spectrum of skills that they learn continuously expands in an organized manner: they undergo a developmental trajectory in which simple skills are learnt first, and skills of progressively increasing complexity are subsequently learnt.

A link can be made to educational systems where research in several domains have tried to study how to provide a good learning experience to learners. This includes the experiences that allow better learning, and in which sequence they must be experienced. This problem is complementary to that of the learner that tries to learn efficiently, and the teacher here has to use as efficiently the limited time and motivational resources of the learner. Several results from psychology [59] and neuroscience [85] have argued that the human brain feels intrinsic pleasure in practicing activities of optimal difficulty or challenge. A teacher must exploit such activities to create positive psychological states of flow [73].

A grand challenge is thus to be able to build machines that possess this capability to discover, adapt and develop continuously new know-how and new knowledge in unknown and changing environments, like human children. In 1950, Turing wrote that the child's brain would show us the way to intelligence: "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's" [154]. Maybe, in opposition to work in the field of Artificial Intelligence who has focused on mechanisms trying to match the capabilities of "intelligent" human adults such as chess playing or natural language

dialogue [91], it is time to take the advice of Turing seriously. This is what a new field, called developmental (or epigenetic) robotics, is trying to achieve [109] [158]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [117], cognitive linguistics [72], and developmental cognitive neuroscience [96] where there has been a considerable amount of research and theories to understand and explain how children learn and develop. A number of general principles are underlying this research agenda: embodiment [63] [131], grounding [89], situatedness [49], self-organization [150] [132], enaction [156], and incremental learning [67].

Among the many issues and challenges of developmental robotics, two of them are of paramount importance: exploration mechanisms and mechanisms for abstracting and making sense of initially unknown sensorimotor channels. Indeed, the typical space of sensorimotor skills that can be encountered and learnt by a developmental robot, as those encountered by human infants, is immensely vast and inhomogeneous. With a sufficiently rich environment and multimodal set of sensors and effectors, the space of possible sensorimotor activities is simply too large to be explored exhaustively in any robot's life time: it is impossible to learn all possible skills and represent all conceivable sensory percepts. Moreover, some skills are very basic to learn, some other very complicated, and many of them require the mastery of others in order to be learnt. For example, learning to manipulate a piano toy requires first to know how to move one's hand to reach the piano and how to touch specific parts of the toy with the fingers. And knowing how to move the hand might require to know how to track it visually.

Exploring such a space of skills randomly is bound to fail or result at best on very inefficient learning [128]. Thus, exploration needs to be organized and guided. The approach of epigenetic robotics is to take inspiration from the mechanisms that allow human infants to be progressively guided, i.e. to develop. There are two broad classes of guiding mechanisms which control exploration:

1. **internal guiding mechanisms,** and in particular intrinsic motivation, responsible of spontaneous exploration and curiosity in humans, which is one of the central mechanisms investigated in FLOWERS, and technically amounts to achieve online active self-regulation of the growth of complexity in learning situations;
2. **social learning and guidance,** a learning mechanisms that exploits the knowledge of other agents in the environment and/or that is guided by those same agents. These mechanisms exist in many different forms like emotional reinforcement, stimulus enhancement, social motivation, guidance, feedback or imitation, some of which being also investigated in FLOWERS;

3.1.1. Internal guiding mechanisms

In infant development, one observes a progressive increase of the complexity of activities with an associated progressive increase of capabilities [117], children do not learn everything at one time: for example, they first learn to roll over, then to crawl and sit, and only when these skills are operational, they begin to learn how to stand. The perceptual system also gradually develops, increasing children perceptual capabilities other time while they engage in activities like throwing or manipulating objects. This make it possible to learn to identify objects in more and more complex situations and to learn more and more of their physical characteristics.

Development is therefore progressive and incremental, and this might be a crucial feature explaining the efficiency with which children explore and learn so fast. Taking inspiration from these observations, some roboticists and researchers in machine learning have argued that learning a given task could be made much easier for a robot if it followed a developmental sequence and "started simple" [53] [79]. However, in these experiments, the developmental sequence was crafted by hand: roboticists manually build simpler versions of a complex task and put the robot successively in versions of the task of increasing complexity. And when they wanted the robot to learn a new task, they had to design a novel reward function.

Thus, there is a need for mechanisms that allow the autonomous control and generation of the developmental trajectory. Psychologists have proposed that intrinsic motivations play a crucial role. Intrinsic motivations are mechanisms that push humans to explore activities or situations that have intermediate/optimal levels of novelty, cognitive dissonance, or challenge [59] [73] [75]. The role and structure of intrinsic motivation in humans have been made more precise thanks to recent discoveries in neuroscience showing the implication

of dopaminergic circuits and in exploration behaviours and curiosity [74] [93] [147]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [128] [129] [145] [57] [94] [112] [146]. While initial models were developed for simple simulated worlds, a current challenge is to manage to build intrinsic motivation systems that can efficiently drive exploratory behaviour in high-dimensional unprepared real world robotic sensorimotor spaces [129], [128], [130], [143]. Specific and complex problems are posed by real sensorimotor spaces, in particular due to the fact that they are both high-dimensional as well as (usually) deeply inhomogeneous. As an example for the latter issue, some regions of real sensorimotor spaces are often unlearnable due to inherent stochasticity or difficulty, in which case heuristics based on the incentive to explore zones of maximal unpredictability or uncertainty, which are often used in the field of active learning [70] [90] typically lead to catastrophic results. The issue of high dimensionality does not only concern motor spaces, but also sensory spaces, leading to the problem of correctly identifying, among typically thousands of quantities, those latent variables that have links to behavioral choices. In FLOWERS, we aim at developing intrinsically motivated exploration mechanisms that scale in those spaces, by studying suitable abstraction processes in conjunction with exploration strategies.

3.1.2. Socially Guided and Interactive Learning

Social guidance is as important as intrinsic motivation in the cognitive development of human babies [117]. There is a vast literature on learning by demonstration in robots where the actions of humans in the environment are recognized and transferred to robots [52]. Most such approaches are completely passive: the human executes actions and the robot learns from the acquired data. Recently, the notion of interactive learning has been introduced in [151], [60], motivated by the various mechanisms that allow humans to socially guide a robot [139]. In an interactive context the steps of self-exploration and social guidance are not separated and a robot learns by self exploration and by receiving extra feedback from the social context [151], [99], [113].

Social guidance is also particularly important for learning to segment and categorize the perceptual space. Indeed, parents interact a lot with infants, for example teaching them to recognize and name objects or characteristics of these objects. Their role is particularly important in directing the infant attention towards objects of interest that will make it possible to simplify at first the perceptual space by pointing out a segment of the environment that can be isolated, named and acted upon. These interactions will then be complemented by the children own experiments on the objects chosen according to intrinsic motivation in order to improve the knowledge of the object, its physical properties and the actions that could be performed with it.

In FLOWERS, we are aiming at including intrinsic motivation system in the self-exploration part thus combining efficient self-learning with social guidance [122], [123]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [110] and robots experiments [95]. Another challenge is to allow for more flexible interaction protocols with the user in terms of what type of feedback is provided and how it is provided [107].

Exploration mechanisms are combined with research in the following directions:

3.1.3. Cumulative learning, reinforcement learning and optimization of autonomous skill learning

FLOWERS develops machine learning algorithms that can allow embodied machines to acquire cumulatively sensorimotor skills. In particular, we develop optimization and reinforcement learning systems which allow robots to discover and learn dictionaries of motor primitives, and then combine them to form higher-level sensorimotor skills.

3.1.4. Autonomous perceptual and representation learning

In order to harness the complexity of perceptual and motor spaces, as well as to pave the way to higher-level cognitive skills, developmental learning requires abstraction mechanisms that can infer structural information out of sets of sensorimotor channels whose semantics is unknown, discovering for example the topology of the body or the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to

be open- ended, progressing in continuous operation from initially simple representations towards abstract concepts and categories similar to those used by humans. Our work focuses on the study of various techniques for:

- autonomous multimodal dimensionality reduction and concept discovery;
- incremental discovery and learning of objects using vision and active exploration, as well as of auditory speech invariants;
- learning of dictionaries of motion primitives with combinatorial structures, in combination with linguistic description;
- active learning of visual descriptors useful for action (e.g. grasping);

3.1.5. Embodiment and maturational constraints

FLOWERS studies how adequate morphologies and materials (i.e. morphological computation), associated to relevant dynamical motor primitives, can importantly simplify the acquisition of apparently very complex skills such as full-body dynamic walking in biped. FLOWERS also studies maturational constraints, which are mechanisms that allow for the progressive and controlled release of new degrees of freedoms in the sensorimotor space of robots.

3.1.6. Discovering and abstracting the structure of sets of uninterpreted sensors and motors

FLOWERS studies mechanisms that allow a robot to infer structural information out of sets of sensorimotor channels whose semantics is unknown, for example the topology of the body and the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to be open-ended, progressing in continuous operation from initially simple representations to abstract concepts and categories similar to those used by humans.

GRAPHDECO Project-Team

3. Research Program

3.1. Introduction

Our research program is oriented around two main axes: 1) Computer-Assisted Design with Heterogeneous Representations and 2) Graphics with Uncertainty and Heterogeneous Content. These two axes are governed by a set of common fundamental goals, share many common methodological tools and are deeply intertwined in the development of applications.

3.1.1. Computer-Assisted Design with Heterogeneous Representations

Designers use a variety of visual representations to explore and communicate about a concept. Fig. 2 illustrates some typical representations, including sketches, hand-made prototypes, 3D models, 3D printed prototypes or instructions.

(a) Ideation sketch (b) Presentation sketch (c) Coarse prototype (d) 3D model

(f) Simulation (e) 3D Printing (g) Technical diagram (h) Instructions

Figure 2. Various representations of a hair dryer at different stages of the design process. Image source, in order: c-maeng on deviantart.com, shauntur on deviantart.com, "Prototyping and Modelmaking for Product Design" Hallgrimsson, B., Laurence King Publishers, 2012, samsher511 on turbosquid.com, my.solidworks.com, weilung tseng on cargocollective.com, howstuffworks.com, u-manual.com

The early representations of a concept, such as rough sketches and hand-made prototypes, help designers formulate their ideas and test the form and function of multiple design alternatives. These low-fidelity representations are meant to be cheap and fast to produce, to allow quick exploration of the *design space* of the concept. These representations are also often approximate to leave room for subjective interpretation and to stimulate imagination; in this sense, these representations can be considered *uncertain*. As the concept gets more finalized, time and effort are invested in the production of more detailed and accurate representations, such as high-fidelity 3D models suitable for simulation and fabrication. These detailed models can also be used to create didactic instructions for assembly and usage.

Producing these different representations of a concept requires specific skills in sketching, modeling, manufacturing and visual communication. For these reasons, professional studios often employ different experts to produce the different representations of the same concept, at the cost of extensive discussions and numerous iterations between the actors of this process. The complexity of the multi-disciplinary skills involved in the design process also hinders their adoption by laymen.

Existing solutions to facilitate design have focused on a subset of the representations used by designers. However, no solution considers all representations at once, for instance to directly convert a series of sketches into a set of physical prototypes. In addition, all existing methods assume that the concept is unique rather than ambiguous. As a result, rich information about the variability of the concept is lost during each conversion step.

We plan to facilitate design for professionals and laymen by adressing the following objectives:

- We want to assist designers in the exploration of the *design space* that captures the possible variations of a concept. By considering a concept as a *distribution* of shapes and functionalities rather than a single object, our goal is to help designers consider multiple design alternatives more quickly and effectively. Such a representation should also allow designers to preserve multiple alternatives along all steps of the design process rather than committing to a single solution early on and pay the price of this decision for all subsequent steps. We expect that preserving alternatives will facilitate communication with engineers, managers and clients, accelerate design iterations and even allow mass personalization by the end consumers.
- We want to support the various representations used by designers during concept development. While drawings and 3D models have received significant attention in past Computer Graphics research, we will also account for the various forms of rough physical prototypes made to evaluate the shape and functionality of a concept. Depending on the task at hand, our algorithms will either analyse these prototypes to generate a virtual concept, or assist the creation of these prototypes from a virtual model. We also want to develop methods capable of adapting to the different drawing and manufacturing techniques used to create sketches and prototypes. We envision design tools that conform to the habits of users rather than impose specific techniques to them.
- We want to make professional design techniques available to novices. Affordable software, hardware and online instructions are democratizing technology and design, allowing small businesses and individuals to compete with large companies. New manufacturing processes and online interfaces also allow customers to participate in the design of an object via mass personalization. However, similarly to what happened for desktop publishing thirty years ago, desktop manufacturing tools need to be simplified to account for the needs and skills of novice designers. We hope to support this trend by adapting the techniques of professionals and by automating the tasks that require significant expertise.

3.1.2. Graphics with Uncertainty and Heterogeneous Content

Our research is motivated by the observation that traditional CG algorithms have not been designed to account for uncertain data. For example, global illumination rendering assumes accurate virtual models of geometry, light and materials to simulate light transport. While these algorithms produce images of high realism, capturing effects such as shadows, reflections and interreflections, they are not applicable to the growing mass of uncertain data available nowadays.

The need to handle uncertainty in CG is timely and pressing, given the large number of *heterogeneous sources of 3D content* that have become available in recent years. These include data from cheap depth+image sensors (e.g., Kinect or the Tango), 3D reconstructions from image/video data, but also data from large 3D geometry databases, or casual 3D models created using simplified sketch-based modeling tools. Such alternate content has varying levels of *uncertainty* about the scene or objects being modelled. This includes uncertainty in geometry, but also in materials and/or lights – which are often not even available with such content. Since CG algorithms cannot be applied directly, visual effects artists spend hundreds of hours correcting inaccuracies and completing the captured data to make them useable in film and advertising.

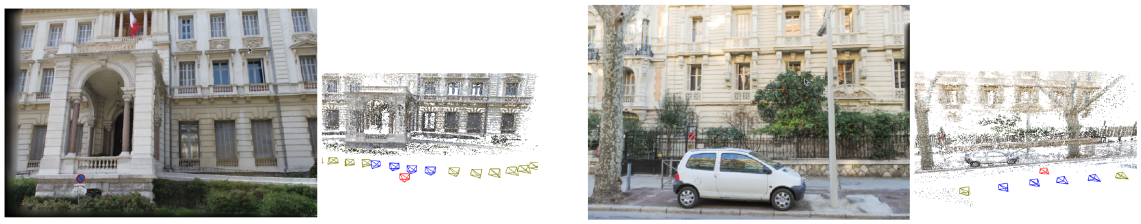


Figure 3. Image-Based Rendering (IBR) techniques use input photographs and approximate 3D to produce new synthetic views.

We identify a major scientific bottleneck which is the need to treat *heterogeneous* content, i.e., containing both (mostly captured) uncertain and perfect, traditional content. Our goal is to provide solutions to this bottleneck, by explicitly and formally modeling uncertainty in CG, and to develop new algorithms that are capable of mixed rendering for this content.

We strive to develop methods in which heterogeneous – and often uncertain – data can be handled automatically in CG with a principled methodology. Our main focus is on *rendering* in CG, including dynamic scenes (video/animations).

Given the above, we need to address the following challenges:

- Develop a theoretical model to handle uncertainty in computer graphics. We must define a new formalism that inherently incorporates uncertainty, and must be able to express traditional CG rendering, both physically accurate and approximate approaches. Most importantly, the new formulation must elegantly handle mixed rendering of perfect synthetic data and captured uncertain content. An important element of this goal is to incorporate *cost* in the choice of algorithm and the optimizations used to obtain results, e.g., preferring solutions which may be slightly less accurate, but cheaper in computation or memory.
- The development of rendering algorithms for heterogeneous content often requires preprocessing of image and video data, which sometimes also includes depth information. An example is the decomposition of images into intrinsic layers of reflectance and lighting, which is required to perform relighting. Such solutions are also useful as image-manipulation or computational photography techniques. The challenge will be to develop such "intermediate" algorithms for the uncertain and heterogeneous data we target.
- Develop efficient rendering algorithms for uncertain and heterogeneous content, reformulating rendering in a probabilistic setting where appropriate. Such methods should allow us to develop approximate rendering algorithms using our formulation in a well-grounded manner. The formalism should include probabilistic models of how the scene, the image and the data interact. These models should be data-driven, e.g., building on the abundance of online geometry and image databases, domain-driven, e.g., based on requirements of the rendering algorithms or perceptually guided, leading to plausible solutions based on limitations of perception.

GRAPHIK Project-Team

3. Research Program

3.1. Logic-based Knowledge Representation and Reasoning

We follow the mainstream *logic-based* approach to knowledge representation (KR). First-order logic (FOL) is the reference logic in KR and most formalisms in this area can be translated into fragments (i.e., particular subsets) of FOL. This is in particular the case for description logics and existential rules, two well-known KR formalisms studied in the team.

A large part of research in this domain can be seen as studying the *trade-off* between the expressivity of languages and the complexity of (sound and complete) reasoning in these languages. The fundamental problem in KR languages is entailment checking: is a given piece of knowledge entailed by other pieces of knowledge, for instance from a knowledge base (KB)? Another important problem is *consistency* checking: is a set of knowledge pieces (for instance the knowledge base itself) consistent, i.e., is it sure that nothing absurd can be entailed from it? The *ontology-mediated query answering* problem is a topical problem (see Section 3.3). It asks for the set of answers to a query in the KB. In the case of Boolean queries (i.e., queries with a yes/no answer), it can be recast as entailment checking.

3.2. Graph-based Knowledge Representation and Reasoning

Besides logical foundations, we are interested in KR formalisms that comply, or aim at complying with the following requirements: to have good *computational* properties and to allow users of knowledge-based systems to have a maximal *understanding and control* over each step of the knowledge base building process and use.

These two requirements are the core motivations for our graph-based approach to KR. We view labelled graphs as an *abstract representation* of knowledge that can be expressed in many KR languages (different kinds of conceptual graphs —historically our main focus— the Semantic Web language RDF (Resource Description Framework), its extension RDFS (RDF Schema), expressive rules equivalent to the so-called tuple-generating-dependencies in databases, some description logics dedicated to query answering, etc.). For these languages, reasoning can be based on the structure of objects, thus based on graph-theoretic notions, while staying logically founded.

More precisely, our basic objects are labelled graphs (or hypergraphs) representing entities and relationships between these entities. These graphs have a natural translation in first-order logic. Our basic reasoning tool is graph homomorphism. The fundamental property is that graph homomorphism is sound and complete with respect to logical entailment *i.e.*, given two (labelled) graphs G and H, there is a homomorphism from G to H *if and only if* the formula assigned to G is entailed by the formula assigned to H. In other words, logical reasoning on these graphs can be performed by graph mechanisms. These knowledge constructs and the associated reasoning mechanisms can be extended (to represent rules for instance) while keeping this fundamental correspondence between graphs and logics.

3.3. Ontology-Mediated Query Answering

Querying knowledge bases has become a central problem in knowledge representation and in databases. A knowledge base (KB) is classically composed of a terminological part (metadata, ontology) and an assertional part (facts, data). Queries are supposed to be at least as expressive as the basic queries in databases, i.e., conjunctive queries, which can be seen as existentially closed conjunctions of atoms or as labelled graphs. The challenge is to define good trade-offs between the expressivity of the ontological language and the complexity of querying data in presence of ontological knowledge. Description logics have been so far the prominent family of formalisms for representing and reasoning with ontological knowledge. However, classical description logics were not designed for efficient data querying. On the other hand, database languages are able to process complex queries on huge databases, but without taking the ontology into account. There is thus a need for new languages and mechanisms, able to cope with the ever growing size of knowledge bases in the Semantic Web or in scientific domains.

This problem is related to two other problems identified as fundamental in KR:

- *Query answering with incomplete information.* Incomplete information means that it might be unknown whether a given assertion is true or false. Databases classically make the so-called closed-world assumption: every fact that cannot be retrieved or inferred from the base is assumed to be false. Knowledge bases classically make the open-world assumption: if something cannot be inferred from the base, and neither can its negation, then its truth status is unknown. The need of coping with incomplete information is a distinctive feature of querying knowledge bases with respect to querying classical databases (however, as explained above, this distinction tends to disappear). The presence of incomplete information makes the query answering task much more difficult.
- *Reasoning with rules.* Researching types of rules and adequate manners to process them is a mainstream topic in the Semantic Web, and, more generally a crucial issue for knowledge-based systems. For several years, we have been studying rules, both in their logical and their graph form, which are syntactically very simple but also very expressive. These rules, known as existential rules or Datalog+, can be seen as an abstraction of ontological knowledge expressed in the main languages used in the context of KB querying.

3.4. Inconsistency and Decision Making

While classical FOL is the kernel of many KR languages, to solve real-world problems we often need to consider features that cannot be expressed purely (or not naturally) in classical logic. The logic and graph-based formalisms used for previous points have thus to be extended with such features. The following requirements have been identified from scenarios in decision making, privileging the agronomy domain:

- to cope with inconsistency;
- to cope with defeasible knowledge;
- to take into account different and potentially conflicting viewpoints;
- to integrate decision notions (priorities, gravity, risk, benefit).

Although the solutions we develop require to be validated on the applications that motivated them, we also want them to be sufficiently generic to be applied in other contexts. One angle of attack (but not the only possible one) consists in increasing the expressivity of our core languages, while trying to preserve their essential combinatorial properties, so that algorithmic optimizations can be transferred to these extensions.

HEPHAISTOS Project-Team

3. Research Program

3.1. Interval analysis

We are interested in real-valued system solving ($f(X) = 0$, $f(X) \leq 0$), in optimization problems, and in the proof of the existence of properties (for example, it exists X such that $f(X) = 0$ or it exist two values X_1, X_2 such that $f(X_1) > 0$ and $f(X_2) < 0$). There are few restrictions on the function f as we are able to manage explicit functions using classical mathematical operators (e.g. $\sin(x + y) + \log(\cos(e^x) + y^2)$ as well as implicit functions (e.g. determining if there are parameter values of a parametrized matrix such that the determinant of the matrix is negative, without calculating the analytical form of the determinant).

Solutions are searched within a finite domain (called a *box*) which may be either continuous or mixed (i.e. for which some variables must belong to a continuous range while other variables may only have values within a discrete set). An important point is that we aim at finding all the solutions within the domain whenever the computer arithmetic will allow it: in other words we are looking for *certified* solutions. For example, for 0-dimensional system solving, we will provide a box that contains one, and only one, solution together with a numerical approximation of this solution. This solution may further be refined at will using multi-precision.

The core of our methods is the use of *interval analysis* that allows one to manipulate mathematical expressions whose unknowns have interval values. A basic component of interval analysis is the *interval evaluation* of an expression. Given an analytical expression F in the unknowns $\{x_1, x_2, ..., x_n\}$ and ranges $\{X_1, X_2, ..., X_n\}$ for these unknowns we are able to compute a range $[A, B]$, called the interval evaluation, such that

$$\forall \{x_1, x_2, ..., x_n\} \in \{X_1, X_2, ..., X_n\}, A \leq F(x_1, x_2, ..., x_n) \leq B \tag{67}$$

In other words the interval evaluation provides a lower bound of the minimum of F and an upper bound of its maximum over the box.

For example if $F = x\ sin(x + x^2)$ and $x \in [0.5, 1.6]$, then $F([0.5, 1.6]) = [-1.362037441, 1.6]$, meaning that for any x in [0.5,1.6] we guarantee that $-1.362037441 \leq f(x) \leq 1.6$.

The interval evaluation of an expression has interesting properties:

- it can be implemented in such a way that the results are guaranteed with respect to round-off errors i.e. property 1 is still valid in spite of numerical errors induced by the use of floating point numbers
- if $A > 0$ or $B < 0$, then no values of the unknowns in their respective ranges can cancel F
- if $A > 0$ ($B < 0$), then F is positive (negative) for any value of the unknowns in their respective ranges

A major drawback of the interval evaluation is that $A(B)$ may be overestimated i.e. values of $x_1, x_2, ..., x_n$ such that $F(x_1, x_2, ..., x_n) = A(B)$ may not exist. This overestimation occurs because in our calculation each occurrence of a variable is considered as an independent variable. Hence if a variable has multiple occurrences, then an overestimation may occur. Such phenomena can be observed in the previous example where $B = 1.6$ while the real maximum of F is approximately 0.9144. The value of B is obtained because we are using in our calculation the formula $F = xsin(y + z^2)$ with y, z having the same interval value as x.

Fortunately there are methods that allow one to reduce the overestimation and the overestimation amount decreases with the width of the ranges. The latter remark leads to the use of a branch-and-bound strategy in which for a given box a variable range will be bisected, thereby creating two new boxes that are stored in a list and processed later on. The algorithm is complete if all boxes in the list have been processed, or if during the process a box generates an answer to the problem at hand (e.g. if we want to prove that $F(X) < 0$, then the algorithm stops as soon as $F(\mathcal{B}) \geq 0$ for a certain box $\mathcal{B}$).

A generic interval analysis algorithm involves the following steps on the current box [8], [4]:

1. *exclusion operators*: these operators determine that there is no solution to the problem within a given box. An important issue here is the extensive and smart use of the monotonicity of the functions
2. *filters*: these operators may reduce the size of the box i.e. decrease the width of the allowed ranges for the variables
3. *existence operators*: they allow one to determine the existence of a unique solution within a given box and are usually associated with a numerical scheme that allows for the computation of this solution in a safe way
4. *bisection*: choose one of the variable and bisect its range for creating two new boxes
5. *storage*: store the new boxes in the list

The scope of the HEPHAISTOS project is to address all these steps in order to find the most efficient procedures. Our efforts focus on mathematical developments (adapting classical theorems to interval analysis, proving interval analysis theorems), the use of symbolic computation and formal proofs (a symbolic pre-processing allows one to automatically adapt the solver to the structure of the problem), software implementation and experimental tests (for validation purposes).

Important note: We have insisted on interval analysis because this is a **major component** or our robotics activity. Our theoretical work in robotics is an analysis of the robotic environment in order to exhibit proofs on the behavior of the system that may be qualitative (e.g. the proof that a cable-driven parallel robot with more than 6 non-deformable cables will have at most 6 cables under tension simultaneously) or quantitative. In the quantitative case as we are dealing with realistic and not toy examples (including our own prototypes that are developed whenever no equivalent hardware is available or to very our assumptions) we have to manage problems that are so complex that analytical solutions are probably out of reach (e.g. the direct kinematics of parallel robots) and we have to resort to algorithms and numerical analysis. We are aware of different approaches in numerical analysis (e.g. some team members were previously involved in teams devoted to computational geometry and algebraic geometry) but interval analysis provides us another approach with high flexibility, the possibility of managing non algebraic problems (e.g. the kinematics of cable-driven parallel robots with sagging cables, that involves inverse hyperbolic functions) and to address various types of issues (system solving, optimization, proof of existence ...). However whenever needed we will rely as well on continuation, algebraic geometry, geometry or learning.

3.2. Robotics

HEPHAISTOS, as a follow-up of COPRIN, has a long-standing tradition of robotics studies, especially for closed-loop robots [3], especially cable-driven parallel robots. We address theoretical issues with the purpose of obtaining analytical and theoretical solutions, but in many cases only numerical solutions can be obtained due to the complexity of the problem. This approach has motivated the use of interval analysis for two reasons:

1. the versatility of interval analysis allows us to address issues (e.g. singularity analysis) that cannot be tackled by any other method due to the size of the problem
2. uncertainties (which are inherent to a robotic device) have to be taken into account so that the *real* robot is guaranteed to have the same properties as the *theoretical* one, even in the worst case. This is a crucial issue for many applications in robotics (e.g. medical or assistance robot)

Our field of study in robotics focuses on *kinematic* issues such as workspace and singularity analysis, positioning accuracy, trajectory planning, reliability, calibration, modularity management and, prominently, *appropriate design*, i.e. determining the dimensioning of a robot mechanical architecture that guarantees that the real robot satisfies a given set of requirements. The methods that we develop can be used for other robotic problems, see for example the management of uncertainties in aircraft design [6].

Our theoretical work must be validated through experiments that are essential for the sake of credibility. A contrario, experiments will feed theoretical work. Hence HEPHAISTOS works with partners on the development of real robots but also develops its own prototypes. In the last years we have developed a large number of prototypes and we have extended our development to devices that are not strictly robots but are part of an overall environment for assistance. We benefit here from the development of new miniature, low energy computers with an interface for analog and logical sensors such as the Arduino or the Phidgets. The web pages http://www-sop.inria.fr/hephaistos/mediatheque/index.html presents all of our prototypes and experimental work.

HYBRID Project-Team

3. Research Program

3.1. Research Program

The scientific objective of Hybrid team is to improve 3D interaction of one or multiple users with virtual environments, by making full use of physical engagement of the body, and by incorporating the mental states by means of brain-computer interfaces. We intend to improve each component of this framework individually, but we also want to improve the subsequent combinations of these components.

The "hybrid" 3D interaction loop between one or multiple users and a virtual environment is depicted in Figure 1 . Different kinds of 3D interaction situations are distinguished (red arrows, bottom): 1) body-based interaction, 2) mind-based interaction, 3) hybrid and/or 4) collaborative interaction (with at least two users). In each case, three scientific challenges arise which correspond to the three successive steps of the 3D interaction loop (blue squares, top): 1) the 3D interaction technique, 2) the modeling and simulation of the 3D scenario, and 3) the design of appropriate sensory feedback.

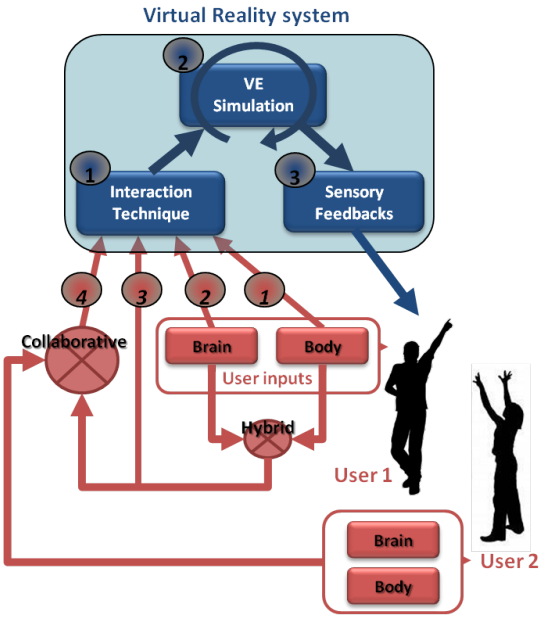


Figure 1. 3D hybrid interaction loop between one or multiple users and a virtual reality system. Top (in blue) three steps of 3D interaction with a virtual environment: (1-blue) interaction technique, (2-blue) simulation of the virtual environment, (3-blue) sensory feedbacks. Bottom (in red) different cases of interaction: (1-red) body-based, (2-red) mind-based, (3-red) hybrid, and (4-red) collaborative 3D interaction.

The 3D interaction loop involves various possible inputs from the user(s) and different kinds of output (or sensory feedback) from the simulated environment. Each user can involve his/her body and mind by means of corporal and/or brain-computer interfaces. A hybrid 3D interaction technique (1) mixes mental and motor inputs and translates them into a command for the virtual environment. The real-time simulation (2) of the

virtual environment is taking into account these commands to change and update the state of the virtual world and virtual objects. The state changes are sent back to the user and perceived by means of different sensory feedbacks (e.g., visual, haptic and/or auditory) (3). The sensory feedbacks are closing the 3D interaction loop. Other users can also interact with the virtual environment using the same procedure, and can eventually "collaborate" by means of "collaborative interactive techniques" (4).

This description is stressing three major challenges which correspond to three mandatory steps when designing 3D interaction with virtual environments:

- **3D interaction techniques:** This first step consists in translating the actions or intentions of the user (inputs) into an explicit command for the virtual environment. In virtual reality, the classical tasks that require such kinds of user command were early categorized in four [44]: navigating the virtual world, selecting a virtual object, manipulating it, or controlling the application (entering text, activating options, etc). The addition of a third dimension, the use of stereoscopic rendering and the use of advanced VR interfaces make however inappropriate many techniques that proved efficient in 2D, and make it necessary to design specific interaction techniques and adapted tools. This challenge is here renewed by the various kinds of 3D interaction which are targeted. In our case, we consider various cases, with motor and/or cerebral inputs, and potentially multiple users.
- **Modeling and simulation of complex 3D scenarios:** This second step corresponds to the update of the state of the virtual environment, in real-time, in response to all the potential commands or actions sent by the user. The complexity of the data and phenomena involved in 3D scenarios is constantly increasing. It corresponds for instance to the multiple states of the entities present in the simulation (rigid, articulated, deformable, fluids, which can constitute both the user's virtual body and the different manipulated objects), and the multiple physical phenomena implied by natural human interactions (squeezing, breaking, melting, etc). The challenge consists here in modeling and simulating these complex 3D scenarios and meeting, at the same time, two strong constraints of virtual reality systems: performance (real-time and interactivity) and genericity (e.g., multi-resolution, multi-modal, multi-platform, etc).
- **Immersive sensory feedbacks:** This third step corresponds to the display of the multiple sensory feedbacks (output) coming from the various VR interfaces. These feedbacks enable the user to perceive the changes occurring in the virtual environment. They are closing the 3D interaction loop, making the user immersed, and potentially generating a subsequent feeling of presence. Among the various VR interfaces which have been developed so far we can stress two kinds of sensory feedback: visual feedback (3D stereoscopic images using projection-based systems such as CAVE systems or Head Mounted Displays); and haptic feedback (related to the sense of touch and to tactile or force-feedback devices). The Hybrid team has a strong expertize in haptic feedback, and in the design of haptic and "pseudo-haptic" rendering [45]. Note that a major trend in the community, which is strongly supported by the Hybrid team, relates to a "perception-based" approach, which aims at designing sensory feedbacks which are well in line with human perceptual capacities.

These three scientific challenges are addressed differently according to the context and the user inputs involved. We propose to consider three different contexts, which correspond to the three different research axes of the Hybrid research team, namely: 1) body-based interaction (motor input only), 2) mind-based interaction (cerebral input only), and then 3) hybrid and collaborative interaction (i.e., the mixing of body and brain inputs from one or multiple users).

3.2. Research Axes

The scientific activity of Hybrid team follows three main axes of research:

- **Body-based interaction in virtual reality.** Our first research axis concerns the design of immersive and effective "body-based" 3D interactions, i.e., relying on a physical engagement of the user's body. This trend is probably the most popular one in VR research at the moment. Most VR setups make use of tracking systems which measure specific positions or actions of the user in order to interact with a virtual environment. However, in recent years, novel options have emerged for measuring

"full-body" movements or other, even less conventional, inputs (e.g. body equilibrium). In this first research axis we are thus concerned by the emergence of new kinds of "body-based interaction" with virtual environments. This implies the design of novel 3D user interfaces and novel 3D interactive techniques, novel simulation models and techniques, and novel sensory feedbacks for body-based interaction with virtual worlds. It involves real-time physical simulation of complex interactive phenomena, and the design of corresponding haptic and pseudo-haptic feedback.

- **Mind-based interaction in virtual reality.** Our second research axis concerns the design of immersive and effective "mind-based" 3D interactions in Virtual Reality. Mind-based interaction with virtual environments is making use of Brain-Computer Interface technology. This technology corresponds to the direct use of brain signals to send "mental commands" to an automated system such as a robot, a prosthesis, or a virtual environment. BCI is a rapidly growing area of research and several impressive prototypes are already available. However, the emergence of such a novel user input is also calling for novel and dedicated 3D user interfaces. This implies to study the extension of the mental vocabulary available for 3D interaction with VE, then the design of specific 3D interaction techniques "driven by the mind" and, last, the design of immersive sensory feedbacks that could help improving the learning of brain control in VR.
- **Hybrid and collaborative 3D interaction.** Our third research axis intends to study the combination of motor and mental inputs in VR, for one or multiple users. This concerns the design of mixed systems, with potentially collaborative scenarios involving multiple users, and thus, multiple bodies and multiple brains sharing the same VE. This research axis therefore involves two interdependent topics: 1) collaborative virtual environments, and 2) hybrid interaction. It should end up with collaborative virtual environments with multiple users, and shared systems with body and mind inputs.

ILDA Project-Team

3. Research Program

3.1. Introduction

Our ability to acquire or generate, store, process, interlink and query data has increased spectacularly over the last few years. The corresponding advances are commonly grouped under the umbrella of so called *Big Data*. Even if the latter has become a buzzword, these advances are real, and they are having a profound impact in domains as varied as scientific research, commerce, social media, industrial processes or e-government. Yet, looking ahead, emerging technologies related to what we now call the *Web of Data* (a.k.a the Semantic Web) have the potential to create an even larger revolution in data-driven activities, by making information accessible to machines as semistructured data [26] that eventually becomes actionable knowledge. Indeed, novel Web data models considerably ease the interlinking of semi-structured data originating from multiple independent sources. They make it possible to associate machine-processable semantics with the data. This in turn means that heterogeneous systems can exchange data, infer new data using reasoning engines, and that software agents can cross data sources, resolving ambiguities and conflicts between them [77]. Datasets are becoming very rich and very large. They are gradually being made even larger and more heterogeneous, but also much more useful, by interlinking them, as exemplified by the Linked Data initiative [49].

These advances raise research questions and technological challenges that span numerous fields of computer science research: databases, communication networks, security and trust, data mining, as well as human-computer interaction. Our research is based on the conviction that interactive systems play a central role in many data-driven activity domains. Indeed, no matter how elaborate the data acquisition, processing and storage pipelines are, data eventually get processed or consumed one way or another by users. The latter are faced with large, increasingly interlinked heterogeneous datasets (see, *e.g.*, Figure 1) that are organized according to complex structures, resulting in overwhelming amounts of both raw data and structured information. Users thus require effective tools to make sense of their data and manipulate them.

Figure 1. Linking Open Data cloud diagram from 2007 to 2017 – http://lod-cloud.net

We approach this problem from the perspective of the Human-Computer Interaction (HCI) field of research, whose goal is to study how humans interact with computers and inspire novel hardware and software designs aimed at optimizing properties such as efficiency, ease of use and learnability, in single-user or cooperative work contexts. More formally, HCI is about designing systems that lower the barrier between users' cognitive model of what they want to accomplish, and computers' understanding of this model. HCI is about the design, implementation and evaluation of computing systems that humans interact with [54], [79]. It is a highly multidisciplinary field, with experts from computer science, cognitive psychology, design, engineering, ethnography, human factors and sociology.

In this broad context, ILDA aims at designing interactive systems that display [35], [61], [87] the data and let users interact with them, aiming to help users better *navigate* and *comprehend* large webs of data represented visually [7], as well as *relate* and *manipulate* them.

Our research agenda consists of the three complementary axes detailed in the following subsections. Designing systems that consider interaction in close conjunction with data semantics is pivotal to all three axes. Those semantics will help drive navigation in, and manipulation of, the data, so as to optimize the communication bandwidth between users and data.

3.2. Semantics-driven Data Manipulation

Participants: Emmanuel Pietriga, Caroline Appert, Anastasia Bezerianos, Marie Destandau, Hugo Romat, Tong Xue, Léo Colombaro.

The Web of Data has been maturing for the last fifteen years and is starting to gain adoption across numerous application domains (Figure 1). Now that most foundational building blocks are in place, from knowledge representation, inference mechanisms and query languages [50], all the way up to the expression of data presentation knowledge [70] and to mechanisms like look-up services [86] or spreading activation [43], we need to pay significant attention to how human beings are going to interact with this new Web, if it is to *"reach its full potential"* [44].

Most efforts in terms of user interface design and development for the Web of data have essentially focused on tools for software developers or subject-matter experts who create ontologies and populate them [56], [41]. Tools more oriented towards end-users are starting to appear [32], [34], [51], [52], [55], [64], including the so-called *linked data browsers* [49]. However, those browsers are in most cases based on quite conventional point-and-click hypertext interfaces that present data to users in a very page-centric, web-of-documents manner that is ill-suited to navigating in, and manipulating, webs of data.

To be successful, interaction paradigms that let users navigate and manipulate data on the Web have to be tailored to the radically different way of browsing information enabled by it, where users directly interact with the data rather than with monolithic documents. The general research question addressed in this part of our research program is how to design novel interaction techniques that help users manipulate their data more efficiently. By data manipulation, we mean all low-level tasks related to manually creating new content, modifying and cleaning existing content, merging data from different sources, establishing connections between datasets, categorizing data, and eventually sharing the end results with other users; tasks that are currently considered quite tedious because of the sheer complexity of the concepts, data models and syntax, and the interplay between all of them.

Our approach is based on the conviction that there is a strong potential for cross-fertilization, as mentioned earlier: on the one hand, user interface design is essential to the management and understanding of webs of data; on the other hand, interlinked datasets enriched with even a small amount of semantics can help create more powerful user interfaces, that provide users with the right information at the right time.

We envision systems that focus on the data themselves, exploiting the underlying *semantics and structure* in the background rather than exposing them – which is what current user interfaces for the Web of Data often do. We envision interactive systems in which the semantics and structure are not exposed directly to users, but serve as input to the system to generate interactive representations that convey information relevant to the task at hand and best afford the possible manipulation actions.

Relevant publications by team members this year: [22], [15], [17] and major ones in recent years: [7].

3.3. Generalized Multi-scale Navigation

Participants: Caroline Appert, Anastasia Bezerianos, Olivier Chapuis, Emmanuel Pietriga, Vanessa Peña-Araya, Marie Destandau, Anna Gogolou, Hugo Romat, Dylan Lebout.

The foundational question addressed here is what to display when, where and how, so as to provide effective support to users in their data understanding and manipulation tasks. ILDA targets contexts in which workers have to interact with complementary views on the same data, or with views on different-but-related datasets, possibly at different levels of abstraction. Being able to combine or switch between representations of the data at different levels of detail and merge data from multiple sources in a single representation is central to many scenarios. This is especially true in both of the application domains we consider: mission-critical systems (e.g., natural disaster crisis management) and the exploratory analysis of scientific data (e.g., correlate theories and heterogeneous observational data for an analysis of a given celestial body in Astrophysics).

A significant part of our research over the last ten years has focused on multi-scale interfaces. We designed and evaluated novel interaction techniques, but also worked actively on the development of open-source UI toolkits for multi-scale interfaces (http://zvtm.sf.net). These interfaces let users navigate large but relatively homogeneous datasets at different levels of detail, on both workstations [73], [29], [69], [68], [67], [30], [72], [28], [74] and wall-sized displays [63], [58], [71], [62], [31], [37], [36]. This part of the ILDA research program is about extending multi-scale navigation in two directions: 1. Enabling the representation of multiple, spatially-registered but widely varying, multi-scale data layers in Geographical Information Systems (GIS); 2. Generalizing the multi-scale navigation paradigm to interconnected, heterogeneous datasets as found on the Web of Data.

The first research problem has been mainly investigated in collaboration with IGN in the context of ANR project MapMuxing, which stands for *multi-dimensional map multiplexing*, from 2014 to early 2019. Project MapMuxing aimed at going beyond the traditional pan & zoom and overview+detail interface schemes, and at designing and evaluating novel cartographic visualizations that rely on high-quality generalization, *i.e.*, the simplification of geographic data to make it legible at a given map scale [82], [83], and symbol specification. Beyond project MapMuxing, we are also investigating multi-scale multiplexing techniques for geo-localized data in the specific context of ultra-high-resolution wall-sized displays, where the combination of a very high pixel density and large physical surface enable us to explore designs that involve collaborative interaction and physical navigation in front of the workspace. This is work done in cooperation with team Massive Data at Inria Chile.

The second research problem is about the extension of multi-scale navigation to interconnected, heterogeneous datasets. Generalization has a rather straightforward definition in the specific domain of geographical information systems, where data items are geographical entities that naturally aggregate as scale increases. But it is unclear how generalization could work for representations of the more heterogeneous webs of data that we consider in the first axis of our research program. Those data form complex networks of resources with multiple and quite varied relationships between them, that cannot rely on a single, unified type of representation (a role played by maps in GIS applications).

Addressing the limits of current generalization processes is a longer-term, more exploratory endeavor. Here again, the machine-processable semantics and structure of the data give us an opportunity to rethink how users navigate interconnected heterogeneous datasets. Using these additional data, we investigate ways to generalize the multi-scale navigation paradigm to datasets whose layout and spatial relationships can be much richer and much more diverse than what can be encoded with static linear hierarchies as typically found today in interfaces for browsing maps or large imagery. Our goal is thus to design and develop highly dynamic and versatile multi-scale information spaces for heterogeneous data whose structure and semantics are not known in advance, but discovered incrementally.

Relevant publications by team members this year: [24], [20], [13], [14], [11], [19] and major ones in recent years: [10], [2].

3.4. Novel Forms of Input for Groups and Individuals

Participants: Caroline Appert, Anastasia Bezerianos, Olivier Chapuis, Emmanuel Pietriga, Eugénie Brasier, Emmanuel Courtoux, Raphaël James.

Analyzing and manipulating large datasets can involve multiple users working together in a coordinated manner in multi-display environments: workstations, handheld devices, wall-sized displays [31]. Those users work towards a common goal, navigating and manipulating data displayed on various hardware surfaces in a coordinated manner. Group awareness [48], [25] is central in these situations, as users, who may or may not be co-located in the same room, can have an optimal individual behavior only if they have a clear picture of what their collaborators have done and are currently doing in the global context. We work on the design and implementation of interactive systems that improve group awareness in co-located situations [57], making individual users able to figure out what other users are doing without breaking the flow of their own actions.

In addition, users need a rich interaction vocabulary to handle large, structured datasets in a flexible and powerful way, regardless of the context of work. Input devices such as mice and trackpads provide a limited number of input actions, thus requiring users to switch between modes to perform different types of data manipulation and navigation actions. The action semantics of these input devices are also often too much dependent on the display output. For instance, a mouse movement and click can only be interpreted according to the graphical controller (widget) above which it is moved. We focus on designing powerful input techniques based upon technologies such as tactile surfaces (supported by UI toolkits developed in-house), 3D motion tracking systems, or custom-built controllers [60] *to complement (rather than replace) traditional input devices* such as keyboards, that remain the best method so far for text entry, and indirect input devices such as mice or trackpads for pixel-precise pointing actions.

The input vocabularies we investigate enable users to navigate and manipulate large and structured datasets in environments that involve multiple users and displays that vary in their size, position and orientation [31], [45], each having their own characteristics and affordances: wall displays [63], [89], workstations, tabletops [66], [40], tablets [65], [84], smartphones [88], [38], [80], [81], and combinations thereof [39], [85], [62], [31].

We aim at designing rich interaction vocabularies that go far beyond what current touch interfaces offer, which rarely exceeds five gestures such as simple slides and pinches. Designing larger gesture vocabularies requires identifying discriminating dimensions (e.g., the presence or absence of anchor points and the distinction between internal and external frames of reference [65]) in order to structure a space of gestures that interface designers can use as a dictionary for choosing a coherent set of controls. These dimensions should be few and simple, so as to provide users with gestures that are easy to memorize and execute. Beyond gesture complexity, the scalability of vocabularies also depends on our ability to design robust gesture recognizers that will allow users to fluidly chain simple gestures that make it possible to interlace navigation and manipulation actions.

We also study how to further extend input vocabularies by combining touch [65], [88], [66] and mid-air gestures [63] with physical objects [53], [78], [60] and classical input devices such as keyboards to enable users to input commands to the system or to involve other users in their workflow (request for help, delegation, communication of personal findings, etc.) [33], [59]. Gestures and objects encode a lot of information in their shape, dynamics and direction, that can be directly interpreted in relation with the user, independently from the display output. Physical objects can also greatly improve coordination among actors for, e.g., handling priorities or assigning specific roles.

Relevant publications by team members this year: [9], [23], [16], [22] and major ones in recent years: [1], [10], [5], [3], [8].

IMAGINE Project-Team

3. Research Program

3.1. Methodology

As already stressed, thinking of future digital modeling technologies as an Expressive Virtual Pen enabling to seamlessly design, refine and convey animated 3D content, leads to revisit models for shapes, motions and stories from a user-centered perspective. More specifically, inspiring from the user-centered interfaces developed in the Human Computer Interaction domain, we introduced the new concept of user-centered graphical models. Ideally, such models should be designed to behave, under any user action, the way a human user would have predicted. In our case, user's actions may include creation gestures such as sketching to draft a shape or direct a motion, deformation gestures such as stretching a shape in space or a motion in time, or copy-paste gestures to transfer some of the features from existing models to other ones. User-centered graphical models need to incorporate knowledge in order to seamlessly generate the appropriate content from such actions. We are using the following methodology to advance towards these goals:

- Develop high-level models for shapes, motion and stories that embed the necessary knowledge to respond as expected to user actions. These models should provide the appropriate handles for conveying the user's intent while embedding procedural methods that seamlessly take care of the appropriate details and constraints.
- Combine these models with expressive design and control tools such as gesture-based control through sketching, sculpting, or acting, towards interactive environments where users can create a new virtual scene, play with it, edit or refine it, and semi-automatically convey it through a video.

3.2. Validation

Validation is a major challenge when developing digital creation tools: there is no ideal result to compare with, in contrast with more standard problems such as reconstructing existing shapes or motions. Therefore, we had to think ahead about our validation strategy: new models for geometry or animation can be validated, as usually done in Computer Graphics, by showing that they solve a problem never tackled before or that they provide a more general or more efficient solution than previous methods. The interaction methods we are developing for content creation and editing rely as much as possible on existing interaction design principles already validated within the HCI community. We also occasionally develop new interaction tools, most often in collaboration with this community, and validate them through user studies. Lastly, we work with expert users from various application domains through our collaborations with professional artists, scientists from other domains, and industrial partners: these expert users validate the use of our new tools compared to their usual pipeline.

LACODAM Project-Team

3. Research Program

3.1. Introduction

The three original research axes of the LACODAM project-team are the following. First, we briefly introduce these axes, as well as their interplay. We then introduce the axis of *Interpretable AI* (Section 3.4), whose emergence is a response to the current societal needs.

- The first research axis (Section 3.2) is dedicated to the design of *novel pattern mining methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. The work on this axis will serve as foundations for work on the other two axes. Thus, this axis will have the strongest impact on our overall goals.
- The second axis (Section 3.3) tackles another aspect of knowledge discovery in data: the *interaction between the user and the system* in order to co-discover novel knowledge. Our team has plenty of experience collaborating with domain experts, and is therefore aware of the need to improve such interaction.
- The third axis (Section 3.4) concerns *decision support*. With the help of methods from the two previous axes, our goal here is to design systems that can either assist humans with making decisions, or make relevant decisions in situations where extremely fast reaction is required.

Figure 1 sums up the detailed work presented in the next few pages: we show the three research axes of the team (X-axis) on the left and our main applications areas (Y-axis) below. In the middle there are colored squares that represent the precise research topics of the team aligned with their axis and main application area. These research topics will be described in this section. Lines represent projects that can link several topics, and that are also connected to their main application area.

3.2. Pattern mining algorithms

Twenty years of research in pattern mining have resulted in efficient approaches to handle the algorithmic complexity of the problem. Existing algorithms are now able to efficiently extract patterns with complex structures (ex: sequences, graphs, co-variations) from large datasets. However, when dealing with large, real-world datasets, these methods still output a huge set of patterns, which is impractical for human analysis. This problem is called *pattern explosion*. The ongoing challenge of pattern mining research is to extract fewer but more meaningful patterns. The LACODAM team is committed to solve the pattern explosion problem by pursuing the following four research topics:

1. the design of dedicated algorithms for mining temporal patterns
2. the design of flexible pattern mining approaches
3. the automatic selection of interesting data mining results
4. the design of parallel pattern algorithms to ensure scalability

The originality of our contributions relies on the exploration of knowledge-based approaches whose principle is to incorporate dedicated domain knowledge (aka application background knowledge) deep into the mining process. While most data mining approaches are based on agnostic approaches designed to cope with pattern explosion, we propose to develop data mining techniques that rely on knowledge-based artificial intelligence techniques. This entails the use of structured knowledge representations, as well as reasoning methods, in combination with mining.

The first topic concerns classical pattern mining in conjunction with expert knowledge in order to define new pattern types (and related algorithms) that can solve applicative issues. In particular, we investigate how to handle temporality in pattern representations which turns out to be important in many real world applications (in particular for decision support) and deserves particular attention.

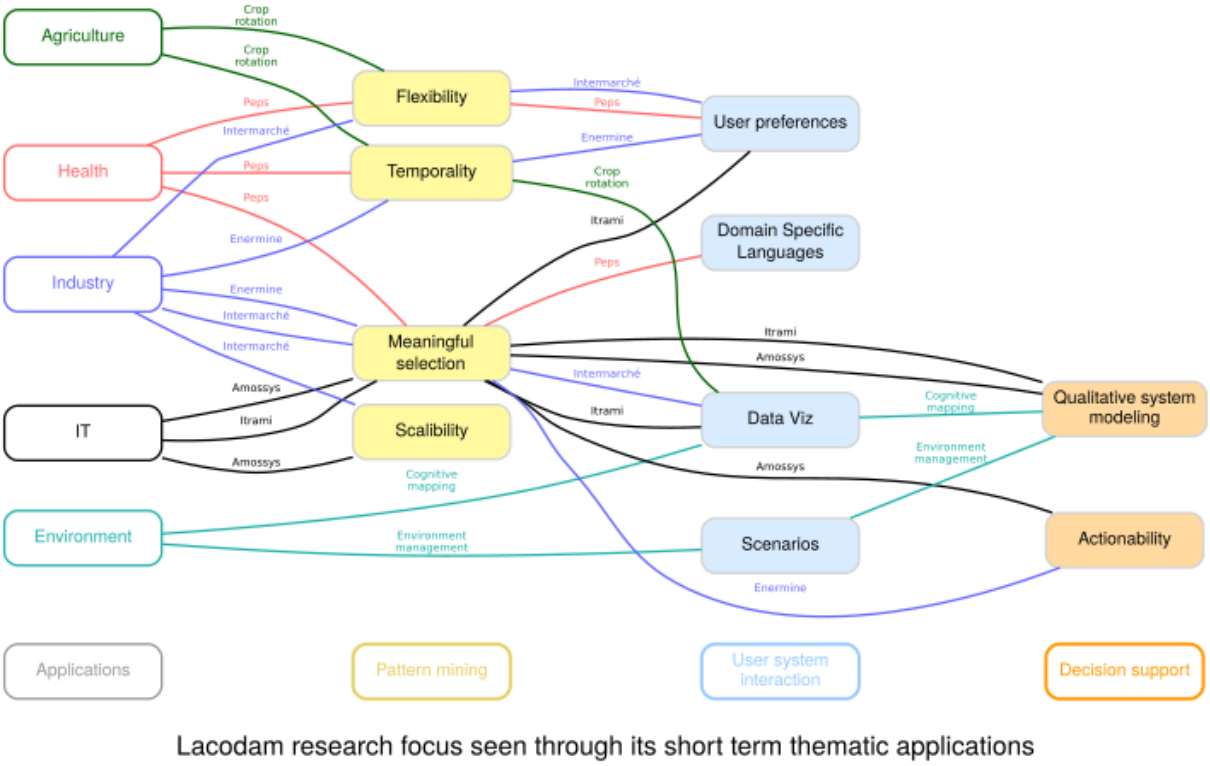


Figure 1. LACODAM research topics organized by axis and application

The next two topics aim at proposing alternative pattern mining methods to let the user incorporate, on her own, knowledge that will help define her pattern domain of interest. Flexible pattern mining approaches enable analysts to easily incorporate extra knowledge, for example domain related constraints, in order to extract only the most relevant patterns. On the other hand, the selection of interesting data mining results aims at devising strategies to filter out the results that are useless to the data analyst. Besides the challenge of algorithmic efficiency, we are interested in formalizing the foundations of interestingness, according to background knowledge modeled with logical knowledge representation paradigms.

Last but not least, pattern mining algorithms are compute-intensive. It is thus important to exploit all the available computing power. Parallelism is for a foreseeable future one of the main ways to speed up computations, and we have a strong competence on the design of parallel pattern mining algorithms. We will exploit this competence in order to guarantee that our approaches scale up to the data provided by our partners.

3.3. User/system interaction

As we pointed out before, there is a strong need to present relevant patterns to the user. This can be done by using more specific constraints, background knowledge and/or tailor-made optimization functions. Due to the difficulty of determining these elements beforehand, one of the most promising solutions is that the system and the user co-construct the definition of relevance, i.e., to have a human in the loop. This requires to have means to present intermediate results to the user, and to get user feedback in order to guide the search space exploration process in the right direction. This is an important research axis for LACODAM, which will be tackled in several complementary ways:

- *Domain Specific Languages:* One way to interact with the user is to propose a Domain Specific Language (DSL) tailored to the domain at hand and to the analysis tasks. The challenge is to propose a DSL allowing the users to easily express the required processing workflows, to deploy those workflows for mining on large volumes of data and to offer as much automation as possible.
- *What if / What for scenarios:* We also investigate the use of scenarios to query results from data mining processes, as well as other complex processes such as complex system simulations or model predictions. Such scenarios are answers to questions of the type “what if [situation]?” or “what [should be done] for [expected outcome]?”.
- *User preferences:* In exploratory analysis, users often do not have a precise idea of what they want, and are not able to formulate such queries. Hence, in LACODAM we investigate simple ways for users to express their interests and preferences, either during the mining process – to guide the search space exploration –, or afterwards during the filtering and interpretation of the most relevant results.
- *Data visualization:* Most of the research directions presented in this document require users to examine patterns at some point. The output of most pattern mining algorithms is usually a (long) list of patterns. While this presentation can be sufficient for some applications, often it does not provide a complete understanding, especially for non-experts in pattern mining. A transversal research topic that we want to explore in LACODAM is to propose data visualization techniques that are adequate for understanding output results. Numerous (failed) experiments have shown that data mining and data visualization are fields, which require distinct skills, thus researchers in one field usually do not make significant advances in the other field (this is detailed in [Keim 2010]). Thus, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4. Decision support

Patterns have proved to be quite useful for decision-aid. Predictive sequential patterns, to give an example, have a direct application in diagnosis. Itemsets and contrast patterns can be used for interpretable machine learning (ML). In regards to diagnosis, LACODAM inherits, from the former DREAM team, a strong background in decision support systems with internationally recognized expertise in this field. This subfield of AI (Artificial Intelligence) is concerned with determining whether a system is operating normally or not, and the cause of

faulty behaviors. The studied system can be an agro- or eco-system, a software system (e.g., a ML classifier), a living being, etc. In relation to interpretable machine learning (ML), this subfield is concerned with the conception of models whose answers are understandable by users. This can be achieved by inducing inherently white-box models from data such as rule-based classifiers/regressors, or by mining rules and explanations from black-box models. The latter setting is quite common due to the high accuracy of black-box models compared to natively interpretable models. Pattern mining is a powerful tool to mine explanations from black-box systems. Those explanations can be used to diagnose biases in systems, either to debug and improve the model, or to generate trust in the verdicts of intelligent software agents.

The increasing volumes of data coming from a range of different systems (ex: sensor data from agro-environmental systems, log data from software systems and ML models, biological data coming from health monitoring systems) can help human and software agents make better decisions. Hence, LACODAM builds upon the idea that decision support systems (an interest bequeathed from DREAM) should take advantage of the available data. This third and last research axis is thus a meeting point for all members of the team, as it requires the integration of AI techniques for traditional decision support systems with results from data mining techniques.

Three main research sub-axes are investigated in LACODAM:

- *Diagnosis-based approaches.* We are exploring how to integrate knowledge found from pattern mining approaches, possibly with the help of interactive methods, into the qualitative models. The goal of such work is to automate as much as possible the construction of prediction models, which can require a lot of human effort.
- *Actionable patterns and rules.* In many settings of "exploratory data mining", the actual interestingness of a pattern is hard to assess, as it may be subjective. However, for some applications there are well defined measures of interestingness and applicability for patterns. Patterns and rules that can lead to actual actions –that are relevant to the user– are called "actionable patterns" and are of vital importance to industrial settings.
- *Mining explanations from ML systems.* Interpretable ML and AI is a current trend for technical, ethical, and legal reasons [27]. In this regard, pattern mining can be used to spot regularities that arise when a complex black-box model yields a particular verdict. For instance, one may want to know the conditions under which the control module of a self-driving car decided to stop without apparent reason, or which factors caused a ML-based credit assessor to reject a loan request. Patterns and conditions are the building blocks for the generation of human-readable explanations for such black-box systems.

3.5. Interpretability

The pervasiveness of complex decision support systems, as well as the general consensus about the societal importance of understanding the rationale embedded in such systems [0], has given momentum to the field of interpretable ML. Being a team specialized in data science, we are fully aware that many problems can be solved by means of complex and accurate ML models. Alas, this accuracy sometimes comes at the expense of interpretability, which can be a major requirement in some contexts (e.g., regression using expertise/rule mining). For this reason, one of the interests of LACODAM is the study of the interpretability-accuracy trade-off. Our studies may be able to answer questions such as "how much accuracy can a model lose (or perhaps gain) by becoming more interpretable?". Such a goal requires us to define interpretability in a more principled way —an endeavour that has been very recently addressed, still not solved. LACODAM is interested in the two main currents of research in interpretability, namely the development of natively interpretable methods, as well as the construction of interpretable mediators between users and black-box models, known as post-hoc interpretability.

[0] General Data Protection Regulation, recital 71 http://www.privacy-regulation.eu/en/r71.htm

We highlight the link between interpretability and LACODAM's axes of decision support, and user/system interaction. In particular, interpretability is a prerequisite for proper user/system interaction and is a central incentive for the advent of data visualization techniques for ML models. This convergence has motivated our interest in *user-oriented post-hoc interpretability*, a sub-field of interpretable ML that adds the user into the formula when generating proper explanations of black-box ML algorithms. This rationale is supported by existing work [28] that suggests that interpretability possesses a subjective component known as plausibility. Moreover, our user-oriented vision meets with the notion of semantic interpretability, where an explanation may resort to high level semantic elements (objects in image classification, or verbal phases in natural language processing) instead of surrogate still-machine-friendly features (such as super-pixels). LACODAM will tackle all these unaddressed aspects of interpretable ML with other Inria teams through the IPL HyAIAI.

3.6. Long-term goals

The following perspectives are at the convergence of the four aforementioned research axes and can be seen as ideal towards our goals:

- *Automating data science workflow discovery.* The current methods for knowledge extraction and construction of decision support systems require a lot of human effort. Our three research axes aim at alleviating this effort, by devising methods that are more generic and by improving the interaction between the user and the system. An ideal solution would be that the user could forget completely about the existence of pattern mining or decision support methods. Instead the user would only loosely specify her problem, while the system constructs various data science / decision support workflows, possibly further refined via interactions.

  We consider that this is a second order AI task, where AI techniques such as planning are used to explore the workflow search space, the workflow itself being composed of data mining and/or decision support components. This is a strategic evolution for data science endeavors, were the demand far exceeds the available human skilled manpower.

- *Logic argumentation based on epistemic interest.* Having increasingly automated approaches will require better and better ways to handle the interactions with the user. Our second long term goal is to explore the use of logic argumentation, i.e., the formalisation of human strategies for reasoning and arguing, in the interaction between users and data analysis tools. Alongside visualization and interactive data mining tools, logic argumentation can be a way for users to query both the results and the way they are obtained. Such querying can also help the expert to reformulate her query in an interactive analysis setting.

  This research direction aims at exploiting principles of interactive data analysis in the context of epistemic interestingness measures. Logic argumentation can be a natural tool for interactions between the user and the system: display of possibly exhaustive list of arguments, relationships between arguments (e.g., reinforcement, compatibility or conflict), possible solutions for argument conflicts, etc.

  The first step is to define a formal argumentation framework for explaining data mining results. This implies to continue theoretical work on the foundations of argumentation in order to identify the most adapted framework (either existing or a new one to be defined). Logic argumentation may be implemented and deeply explored in ASP, allowing us to build on our expertise in this logic language.

- *Collaborative feedback and knowledge management.* We are convinced that improving the data science process, and possibly automating it, will rely on high-quality feedback from communities on the web. Consider for example what has been achieved by collaborative platforms such as StackOverflow: it has become the reference site for any programming question.

  Data science is a more complex problem than programming, as in order to get help from the community, the user has to share her data and workflow, or at least some parts of them. This raises obvious privacy issues that may prevent this idea to succeed. As our research on automating the

production of data science workflows should enable more people to have access to data science results, we are interested in the design of collaborative platforms to exchange expert advices over data, workflows and analysis results. This aims at exploiting human feedback to improve the automation of data science system via machine learning methods.

LARSEN Project-Team

3. Research Program

3.1. Lifelong Autonomy

3.1.1. Scientific Context

So far, only a few autonomous robots have been deployed for a long time (weeks, months, or years) outside of factories and laboratories. They are mostly mobile robots that simply "move around" (e.g., vacuum cleaners or museum "guides") and data collecting robots (e.g., boats or underwater "gliders" that collect data about the water of the ocean).

A large part of the long-term autonomy community is focused on simultaneous localization and mapping (SLAM), with a recent emphasis on changing and outdoor environments [25], [34]. A more recent theme is life-long learning: during long-term deployment, we cannot hope to equip robots with everything they need to know, therefore some things will have to be learned along the way. Most of the work on this topic leverages machine learning and/or evolutionary algorithms to improve the ability of robots to react to unforeseen changes [25], [32].

3.1.2. Main Challenges

The first major challenge is to endow robots with a stable situation awareness in open and dynamic environments. This covers both the state estimation of the robot itself as well as the perception/representation of the environment. Both problems have been claimed to be solved but it is only the case for static environments [30].

In the LARSEN team, we aim at deployment in environments shared with humans which imply dynamic objects that degrade both the mapping and localization of a robot, especially in cluttered spaces. Moreover, when robots stay longer in the environment than for the acquisition of a snapshot map, they have to face structural changes, such as the displacement of a piece of furniture or the opening or closing of a door. The current approach is to simply update an implicitly static map with all observations with no attempt at distinguishing the suitable changes. For localization in not-too-cluttered or not-too-empty environments, this is generally sufficient as a significant fraction of the environment should remain stable. But for life-long autonomy, and in particular navigation, the quality of the map, and especially the knowledge of the stable parts, is primordial.

A second major obstacle to move robots outside of labs and factories is their fragility: Current robots often break in a few hours, if not a few minutes. This fragility mainly stems from the overall complexity of robotic systems, which involve many actuators, many sensors, and complex decisions, and from the diversity of situations that robots can encounter. Low-cost robots exacerbate this issue because they can be broken in many ways (high-quality material is expensive), because they have low self-sensing abilities (sensors are expensive and increase the overall complexity), and because they are typically targeted towards non-controlled environments (e.g., houses rather than factories, in which robots are protected from most unexpected events). More generally, this fragility is a symptom of the lack of adaptive abilities in current robots.

3.1.3. Angle of Attack

To solve the state estimation problem, our approach is to combine classical estimation filters (Extended Kalman Filters, Unscented Kalman Filters, or particle filters) with a Bayesian reasoning model in order to internally simulate various configurations of the robot in its environment. This should allow for adaptive estimation that can be used as one aspect of long-term adaptation. To handle dynamic and structural changes in an environment, we aim at assessing, for each piece of observation, whether it is static or not.

We also plan to address active sensing to improve the situation awareness of robots. Literally, active sensing is the ability of an interacting agent to act so as to control what it senses from its environment with the typical objective of acquiring information about this environment. A formalism for representing and solving active sensing problems has already been proposed by members of the team [24] and we aim to use this to formalize decision making problems of improving situation awareness.

Situation awareness of robots can also be tackled by cooperation, whether it be between robots or between robots and sensors in the environment (led out intelligent spaces) or between robots and humans. This is in rupture with classical robotics, in which robots are conceived as self-contained. But, in order to cope with as diverse environments as possible, these classical robots use precise, expensive, and specialized sensors, whose cost prohibits their use in large-scale deployments for service or assistance applications. Furthermore, when all sensors are on the robot, they share the same point of view on the environment, which is a limit for perception. Therefore, we propose to complement a cheaper robot with sensors distributed in a target environment. This is an emerging research direction that shares some of the problematics of multi-robot operation and we are therefore collaborating with other teams at Inria that address the issue of communication and interoperability.

To address the fragility problem, the traditional approach is to first diagnose the situation, then use a planning algorithm to create/select a contingency plan. But, again, this calls for both expensive sensors on the robot for the diagnosis and extensive work to predict and plan for all the possible faults that, in an open and dynamic environment, are almost infinite. An alternative approach is then to skip the diagnosis and let the robot discover by trial and error a behavior that works in spite of the damage with a reinforcement learning algorithm [39], [32]. However, current reinforcement learning algorithms require hundreds of trials/episodes to learn a single, often simplified, task [32], which makes them impossible to use for real robots and more ambitious tasks. We therefore need to design new trial-and-error algorithms that will allow robots to learn with a much smaller number of trials (typically, a dozen). We think the key idea is to guide online learning on the physical robot with dynamic simulations. For instance, in our recent work, we successfully mixed evolutionary search in simulation, physical tests on the robot, and machine learning to allow a robot to recover from physical damage [33], [1].

A final approach to address fragility is to deploy several robots or a swarm of robots or to make robots evolve in an active environment. We will consider several paradigms such as (1) those inspired from collective natural phenomena in which the environment plays an active role for coordinating the activity of a huge number of biological entities such as ants and (2) those based on online learning [29]. We envision to transfer our knowledge of such phenomenon to engineer new artificial devices such as an intelligent floor (which is in fact a spatially distributed network in which each node can sense, compute and communicate with contiguous nodes and can interact with moving entities on top of it) in order to assist people and robots (see the principle in [37], [29], [23]).

3.2. Natural Interaction with Robotic Systems

3.2.1. Scientific Context

Interaction with the environment is a primordial requirement for an autonomous robot. When the environment is sensorized, the interaction can include localizing, tracking, and recognizing the behavior of robots and humans. One specific issue lies in the lack of predictive models for human behavior and a critical constraint arises from the incomplete knowledge of the environment and the other agents.

On the other hand, when working in the proximity of or directly with humans, robots must be capable of safely interacting with them, which calls upon a mixture of physical and social skills. Currently, robot operators are usually trained and specialized but potential end-users of robots for service or personal assistance are not skilled robotics experts, which means that the robot needs to be accepted as reliable, trustworthy and efficient [42]. Most Human-Robot Interaction (HRI) studies focus on verbal communication [38] but applications such as assistance robotics require a deeper knowledge of the intertwined exchange of social and physical signals to provide suitable robot controllers.

3.2.2. Main Challenges

We are here interested in building the bricks for a situated Human-Robot Interaction (HRI) addressing both the physical and social dimension of the close interaction, and the cognitive aspects related to the analysis and interpretation of human movement and activity.

The combination of physical and social signals into robot control is a crucial investigation for assistance robots [40] and robotic co-workers [36]. A major obstacle is the control of physical interaction (precisely, the control of contact forces) between the robot and the human while both partners are moving. In mobile robots, this problem is usually addressed by planning the robot movement taking into account the human as an obstacle or as a target, then delegating the execution of this "high-level" motion to whole-body controllers, where a mixture of weighted tasks is used to account for the robot balance, constraints, and desired end-effector trajectories [26].

The first challenge is to make these controllers easier to deploy in real robotics systems, as currently they require a lot of tuning and can become very complex to handle the interaction with unknown dynamical systems such as humans. Here, the key is to combine machine learning techniques with such controllers.

The second challenge is to make the robot react and adapt online to the human feedback, exploiting the whole set of measurable verbal and non-verbal signals that humans naturally produce during a physical or social interaction. Technically, this means finding the optimal policy that adapts the robot controllers online, taking into account feedback from the human. Here, we need to carefully identify the significant feedback signals or some metrics of human feedback. In real-world conditions (i.e., outside the research laboratory environment) the set of signals is technologically limited by the robot's and environmental sensors and the onboard processing capabilities.

The third challenge is for a robot to be able to identify and track people on board. The motivation is to be able to estimate online either the position, the posture, or even moods and intentions of persons surrounding the robot. The main challenge is to be able to do that online, in real-time and in cluttered environments.

3.2.3. Angle of Attack

Our key idea is to exploit the physical and social signals produced by the human during the interaction with the robot and the environment in controlled conditions, to learn simple models of human behavior and consequently to use these models to optimize the robot movements and actions. In a first phase, we will exploit human physical signals (e.g., posture and force measurements) to identify the elementary posture tasks during balance and physical interaction. The identified model will be used to optimize the robot whole-body control as prior knowledge to improve both the robot balance and the control of the interaction forces. Technically, we will combine weighted and prioritized controllers with stochastic optimization techniques. To adapt online the control of physical interaction and make it possible with human partners that are not robotics experts, we will exploit verbal and non-verbal signals (e.g., gaze, touch, prosody). The idea here is to estimate online from these signals the human intent along with some inter-individual factors that the robot can exploit to adapt its behavior, maximizing the engagement and acceptability during the interaction.

Another promising approach already investigated in the LARSEN team is the capability for a robot and/or an intelligent space to localize humans in its surrounding environment and to understand their activities. This is an important issue to handle both for safe and efficient human-robot interaction.

Simultaneous Tracking and Activity Recognition (STAR) [41] is an approach we want to develop. The activity of a person is highly correlated with his position, and this approach aims at combining tracking and activity recognition to benefit one from another. By tracking the individual, the system may help infer its possible activity, while by estimating the activity of the individual, the system may make a better prediction of his/her possible future positions (especially in the case of occlusions). This direction has been tested with simulator and particle filters [28], and one promising direction would be to couple STAR with decision making formalisms like partially observable Markov decision processes (POMDPs). This would allow us to formalize problems such as deciding which action to take given an estimate of the human location and activity. This could also formalize other problems linked to the active sensing direction of the team: how the robotic system

should choose its actions in order to have a better estimate of the human location and activity (for instance by moving in the environment or by changing the orientation of its cameras)?

Another issue we want to address is robotic human body pose estimation. Human body pose estimation consists of tracking body parts by analyzing a sequence of input images from single or multiple cameras.

Human posture analysis is of high value for human robot interaction and activity recognition. However, even if the arrival of new sensors like RGB-D cameras has simplified the problem, it still poses a great challenge, especially if we want to do it online, on a robot and in realistic world conditions (cluttered environment). This is even more difficult for a robot to bring together different capabilities both at the perception and navigation level [27]. This will be tackled through different techniques, going from Bayesian state estimation (particle filtering), to learning, active and distributed sensing.

LINKMEDIA Project-Team

3. Research Program

3.1. Scientific background

LINKMEDIA is de facto a multidisciplinary research team in order to gather the multiple skills needed to enable humans to gain insight into extremely large collections of multimedia material. It is *multimedia data* which is at the core of the team and which drives the design of our scientific contributions, backed-up with solid experimental validations. *Multimedia data*, again, is the rationale for selecting problems, applicative fields and partners.

Our activities therefore include studying the following scientific fields:

- multimedia: content-based analysis; multimodal processing and fusion; multimedia applications;
- computer vision: compact description of images; object and event detection;
- machine learning: deep architectures; structured learning; adversarial learning;
- natural language processing: topic segmentation; information extraction;
- information retrieval: high-dimensional indexing; approximate k-nn search; embeddings;
- data mining: time series mining; knowledge extraction.

3.2. Workplan

Overall, LINKMEDIA follows two main directions of research that are (i) extracting and representing information from the documents in collections, from the relationships between the documents and from what user build from these documents, and (ii) facilitating the access to documents and to the information that has been elaborated from their processing.

3.2.1. Research Direction 1: Extracting and Representing Information

LINKMEDIA follows several research tracks for *extracting* knowledge from the collections and *representing* that knowledge to facilitate users acquiring gradual, long term, constructive insights. Automatically processing documents makes it crucial to consider the accountability of the algorithms, as well as understanding when and why algorithms make errors, and possibly invent techniques that compensate or reduce the impact of errors. It also includes dealing with malicious adversaries carefully manipulating the data in order to compromise the whole knowledge extraction effort. In other words, LINKMEDIA also investigates various aspects related to the *security* of the algorithms analyzing multimedia material for knowledge extraction and representation.

Knowledge is not solely extracted by algorithms, but also by humans as they gradually get insight. This human knowledge can be materialized in computer-friendly formats, allowing algorithms to use this knowledge. For example, humans can create or update ontologies and knowledge bases that are in relation with a particular collection, they can manually label specific data samples to facilitate their disambiguation, they can manually correct errors, etc. In turn, knowledge provided by humans may help algorithms to then better process the data collections, which provides higher quality knowledge to humans, which in turn can provide some better feedback to the system, and so on. This virtuous cycle where algorithms and humans cooperate in order to make the most of multimedia collections requires specific support and techniques, as detailed below.

3.2.1.1. Machine Learning for Multimedia Material.

Many approaches are used to extract relevant information from multimedia material, ranging from very low-level to higher-level descriptions (classes, captions, ...). That diversity of information is produced by algorithms that have varying degrees of supervision. Lately, fully supervised approaches based on deep learning proved to outperform most older techniques. This is particularly true for the latest developments of Recurrent Neural Networkds (RNN, such as LSTMs) or convolutional neural network (CNNs) for images that reach excellent performance [62]. LINKMEDIA contributes to advancing the state of the art in computing representations for multimedia material by investigating the topics listed below. Some of them go beyond the very processing of multimedia material as they also question the fundamentals of machine learning procedures when applied to multimedia.

- *Learning from few samples/weak supervisions.* CNNs and RNNs need large collections of carefully annotated data. They are not fitted for analyzing datasets where few examples per category are available or only cheap image-level labels are provided. LINKMEDIA investigates low-shot, semi-supervised and weakly supervised learning processes: Augmenting scarce training data by automatically propagating labels [65], or transferring what was learned on few very well annotated samples to allow the precise processing of poorly annotated data [74]. Note that this context also applies to the processing of heritage collections (paintings, illuminated manuscripts, ...) that strongly differ from contemporary natural images. Not only annotations are scarce, but the learning processes must cope with material departing from what standard CNNs deal with, as classes such as "planes", "cars", etc, are irrelevant in this case.
- *Ubiquitous Training.* NN (CNNs, LSTMs) are mainstream for producing representations suited for high-quality classification. Their training phase is ubiquitous because the same representations can be used for tasks that go beyond classification, such as retrieval, few-shot, meta- and incremental learning, all boiling down to some form of metric learning. We demonstrated that this ubiquitous training is relatively simpler [65] yet as powerful as ad-hoc strategies fitting specific tasks [79]. We study the properties and the limitations of this ubiquitous training by casting metric learning as a classification problem.
- *Beyond static learning*. Multimedia collections are by nature continuously growing, and ML processes must adapt. It is not conceivable to re-train a full new model at every change, but rather to support continuous training and/or allowing categories to evolve as the time goes by. New classes may be defined from only very few samples, which links this need for dynamicity to the low-shot learning problem discussed here. Furthermore, active learning strategies determining which is the next sample to use to best improve classification must be considered to alleviate the annotation cost and the re-training process [69]. Eventually, the learning process may need to manage an extremely large number of classes, up to millions. In this case, there is a unique opportunity of blending the expertise of LINKMEDIA on large scale indexing and retrieval with deep learning. Base classes can either be "summarized" e.g. as a multi-modal distribution, or their entire training set can be made accessible as an external associative memory [86].
- *Learning and lightweight architectures.* Multimedia is everywhere, it can be captured and processed on the mobile devices of users. It is necessary to study the design of lightweight ML architectures for mobile and embedded vision applications. Inspired by [90], we study the savings from quantizing hyper-parameters, pruning connections or other approximations, observing the trade-off between the footprint of the learning and the quality of the inference. Once strategy of choice is progressive learning which early aborts when confident enough [70].
- *Multimodal embeddings.* We pursue pioneering work of LINKMEDIA on multimodal embedding, i.e., representing multiple modalities or information sources in a single embedded space [83], [85], [84]. Two main directions are explored: exploiting adversarial architectures (GANs) for embedding via translation from one modality to another, extending initial work in [84] to highly heterogeneous content; combining and constraining word and RDF graph embeddings to facilitate entity linking and explanation of lexical co-occurrences [81].

- *Accountability of ML processes.* ML processes achieve excellent results but it is mandatory to verify that accuracy results from having determined an adequate problem representation, and not from being abused by artifacts in the data. LINKMEDIA designs procedures for at least explaining and possibly interpreting and understanding what the models have learned. We consider heat-maps materializing which input (pixels, words) have the most importance in the decisions [77], Taylor decompositions to observe the individual contributions of each relevance scores or estimating LID [47] as a surrogate for accounting for the smoothness of the space.
- *Extracting information.* ML is good at extracting features from multimedia material, facilitating subsequent classification, indexing, or mining procedures. LINKMEDIA designs extraction processes for identifying parts in the images [75], [76], relationships between the various objects that are represented in images [53], learning to localizing objects in images with only weak, image-level supervision [78] or fine-grained semantic information in texts [58]. One technique of choice is to rely on generative adversarial networks (GAN) for learning low-level representations. These representations can e.g. be based on the analysis of density [89], shading, albedo, depth, etc.
- *Learning representations for time evolving multimedia material.* Video and audio are time evolving material, and processing them requests to take their time line into account. In [71], [57] we demonstrated how shapelets can be used to transform time series into time-free high-dimensional vectors, preserving however similarities between time series. Representing time series in a metric space improves clustering, retrieval, indexing, metric learning, semi-supervised learning and many other machine learning related tasks. Research directions include adding localization information to the shapelets, fine-tuning them to best fit the task in which they are used as well as designing hierarchical representations.

3.2.1.2. Adversarial Machine Learning.

Systems based on ML take more and more decisions on our behalf, and maliciously influencing these decisions by crafting adversarial multimedia material is a potential source of dangers: a small amount of carefully crafted noise imperceptibly added to images corrupts classification and/or recognition. This can naturally impact the insight users get on the multimedia collection they work with, leading to taking erroneous decisions e.g.

This adversarial phenomenon is not particular to deep learning, and can be observed even when using other ML approaches [52]. Furthermore, it has been demonstrated that adversarial samples generalize very well across classifiers, architectures, training sets. The reasons explaining why such tiny content modifications succeed in producing severe errors are still not well understood.

We are left with little choice: we must gain a better understanding of the weaknesses of ML processes, and in particular of deep learning. We must understand why attacks are possible as well as discover mechanisms protecting ML against adversarial attacks (with a special emphasis on convolutional neural networks). Some initial contributions have started exploring such research directions, mainly focusing on images and computer vision problems. Very little has been done for understanding adversarial ML from a *multimedia* perspective [56].

LINKMEDIA is in a unique position to throw at this problem new perspectives, by experimenting with other modalities, used in isolation one another, as well as experimenting with true multimodal inputs. This is very challenging, and far more complicated and interesting than just observing adversarial ML from a computer vision perspective. No one clearly knows what is at stake with adversarial audio samples, adversarial video sequences, adversarial ASR, adversarial NLP, adversarial OCR, all this being often part of a sophisticated multimedia processing pipeline.

Our ambition is to lead the way for initiating investigations where the full diversity of modalities we are used to work with in multimedia are considered from a perspective of adversarial attacks and defenses, both at learning and test time. In addition to what is described above, and in order to trust the multimedia material we analyze and/or the algorithms that are at play, LINKMEDIA investigates the following topics:

- *Beyond classification.* Most contributions in relation with adversarial ML focus on classification tasks. We started investigating the impact of adversarial techniques on more diverse tasks such as

retrieval [46]. This problem is related to the very nature of euclidean spaces where distances and neighborhoods can all be altered. Designing defensive mechanisms is a natural companion work.

- *Detecting false information.* We carry-on with earlier pioneering work of LINKMEDIA on false information detection in social media. Unlike traditional approaches in image forensics [60], we build on our expertise in content-based information retrieval to take advantage of the contextual information available in databases or on the web to identify out-of-context use of text or images which contributed to creating a false information [72].
- *Deep fakes.* Progress in deep ML and GANs allow systems to generate realistic images and are able to craft audio and video of existing people saying or doing things they never said or did [68]. Gaining in sophistication, these machine learning-based "deep fakes" will eventually be almost indistinguishable from real documents, making their detection/rebutting very hard. LINKMEDIA develops deep learning based counter-measures to identify such modern forgeries. We also carry on with making use of external data in a provenance filtering perspective [91] in order to debunk such deep fakes.
- *Distributions, frontiers, smoothness, outliers.* Many factors that can possibly explain the adversarial nature of some samples are in relation with their distribution in space which strongly differs from the distribution of natural, genuine, non adversarial samples. We are investigating the use of various information theoretical tools that facilitate observing distributions, how they differ, how far adversarial samples are from benign manifolds, how smooth is the feature space, etc. In addition, we are designing original adversarial attacks and develop detection and curating mechanisms [47].

3.2.1.3. Multimedia Knowledge Extraction.

Information obtained from collections via computer ran processes is not the only thing that needs to be represented. Humans are in the loop, and they gradually improve their level of understanding of the content and nature of the multimedia collection. Discovering knowledge and getting insight is involving multiple people across a long period of time, and what each understands, concludes and discovers must be recorded and made available to others. Collaboratively inspecting collections is crucial. Ontologies are an often preferred mechanism for modeling what is inside a collection, but this is probably limitative and narrow.

LINKMEDIA is concerned with making use of existing strategies in relation with ontologies and knowledge bases. In addition, LINKMEDIA uses mechanisms allowing to materialize the knowledge gradually acquired by humans and that might be subsequently used either by other humans or by computers in order to better and more precisely analyze collections. This line of work is instantiated at the core of the iCODA project LINKMEDIA coordinates. We are therefore concerned with:

- *Multimedia analysis and ontologies.* We develop approaches for linking multimedia content to entities in ontologies for text and images, building on results in multimodal embedding to cast entity linking into a nearest neighbor search problem in a high-dimensional joint embedding of content and entities [85]. We also investigate the use of ontological knowledge to facilitate information extraction from content [9].
- *Explainability and accountability in information extraction.* In relation with ontologies and entity linking, we develop innovative approaches to explain statistical relations found in data, in particular lexical or entity co-occurrences in textual data, for example using embeddings constrained with translation properties of RDF knowledge or path-based explanation within RDF graphs. We also work on confidence measures in entity linking and information extraction, studying how the notions of confidence and information source can be accounted for in knowledge basis and used in human-centric collaborative exploration of collections.
- *Dynamic evolution of models for information extraction.* In interactive exploration and information extraction, e.g., on cultural or educational material, knowledge progressively evolves as the process goes on, requiring on-the-fly design of new models for content-based information extractors from very few examples, as well as continuous adaptation of the models. Combining in a seamless way low-shot, active and incremental learning techniques is a key issue that we investigate to enable this dynamic mechanisms on selected applications.

3.2.1.4. Research Direction 2: Accessing Information

LINKMEDIA centers its activities on enabling humans to make good use of vast multimedia collections. This material takes all its cultural and economic value, all its artistic wonder when it can be accessed, watched, searched, browsed, visualized, summarized, classified, shared, ... This allows users to fully enjoy the incalculable richness of the collections. It also makes it possible for companies to create business rooted in this multimedia material.

Accessing the multimedia data that is inside a collection is complicated by the various type of data, their volume, their length, etc. But it is even more complicated to access the information that is not materialized in documents, such as the relationships between parts of different documents that however share some similarity. LINKMEDIA in its first four years of existence established itself as one of the leading teams in the field of multimedia analytics, contributing to the establishment of a dedicated community (refer to the various special sessions we organized with MMM, the iCODA and the LIMAH projects, as well as [66], [67], [63]).

Overall, facilitating the access to the multimedia material, to the relevant information and the corresponding knowledge asks for algorithms that efficiently *search* collections in order to identify the elements of collections or of the acquired knowledge that are matching a query, or that efficiently allow *navigating* the collections or the acquired knowledge. Navigation is likely facilitated if techniques are able to handle information and knowledge according to hierarchical perspectives, that is, allow to reveal data according to various levels of details. Aggregating or *summarizing* multimedia elements is not trivial.

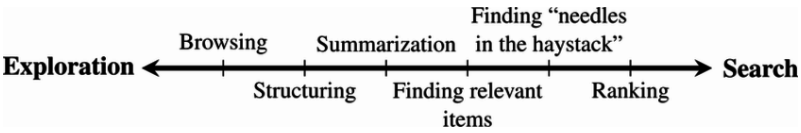


Figure 1. Exploration-search axis with example tasks

Three topics are therefore in relation with this second research direction. LINKMEDIA tackles the issues in relation to searching, to navigating and to summarizing multimedia information. Information needs when discovering the content of a multimedia collection can be conveniently mapped to the exploration-search axis, as first proposed by Zahálka and Worring in [88], and illustrated by Figure 1 where expert users typically work near the right end because their tasks involve precise queries probing search engines. In contrast, lay-users start near the exploration end of the axis. Overall, users may alternate searches and explorations by going back and forth along the axis. The underlying model and system must therefore be highly dynamic, support interactions with the users and propose means for easy refinements. LINKMEDIA contributes to advancing

the state of the art in searching operations, in navigating operations (also referred to as browsing), and in summarizing operations.

3.2.1.4.1. Searching.

Search engines must run similarity searches very efficiently. High-dimensional indexing techniques therefore play a central role. Yet, recent contributions in ML suggest to revisit indexing in order to adapt to the specific properties of modern features describing contents.

- *Advanced scalable indexing.* High-dimensional indexing is one of the foundations of LINKMEDIA. Modern features extracted from the multimedia material with the most recent ML techniques shall be indexed as well. This, however, poses a series of difficulties due to the dimensionality of these features, their possible sparsity, the complex metrics in use, the task in which they are involved (instance search, k-nn, class prototype identification, manifold search [65], time series retrieval, ...). Furthermore, truly large datasets require involving sketching [50], secondary storage and/or distribution [49], [48], alleviating the explosion of the number of features to consider due to their local nature or other innovative methods [64], all introducing complexities. Last, indexing multimodal embedded spaces poses a new series of challenges.
- *Improving quality.* Scalable indexing techniques are approximate, and what they return typically includes a fair amount of false positives. LINKMEDIA works on improving the quality of the results returned by indexing techniques. Approaches taking into account neighborhoods [59], manifold structures instead of pure distance based similarities [65] must be extended to cope with advanced indexing in order to enhance quality. This includes feature selection based on intrinsic dimensionality estimation [47].
- *Dynamic indexing.* Feature collections grow, and it is not an option to fully reindex from scratch an updated collection. This trivially applies to the features directly extracted from the media items, but also to the base class prototypes that can evolve due to the non-static nature of learning processes. LINKMEDIA will continue investigating what is at stake when designing dynamic indexing strategies.

3.2.1.4.2. Navigating.

Navigating a multimedia collection is very central to its understanding. It differs from searching as navigation is not driven by any specific query. Rather, it is mostly driven by the relationships that various documents have one another. Relationships are supported by the links between documents and/or parts of documents. Links rely on semantic similarity, depicting the fact that two documents share information on the same topic. But other aspects than semantics are also at stake, e.g., time with the dates of creation of the documents or geography with mentions or appearance in documents of some geographical landmarks or with geo-tagged data.

In multimedia collections, links can be either implicit or explicit, the latter being much easier to use for navigation. An example of an implicit link can be the name of someone existing in several different news articles; we, as humans, create a mental link between them. In some cases, the computer misses such configurations, leaving such links implicit. Implicit links are subject to human interpretation, hence they are sometimes hard to identify for any automatic analysis process. Implicit links not being materialized, they can therefore hardly be used for navigation or faceted search. Explicit links can typically be seen as hyperlinks, established either by content providers or, more aligned with LINKMEDIA, automatically determined from content analysis. Entity linking (linking content to an entity referenced in a knowledge base) is a good example of the creation of explicit links. Semantic similarity links, as investigated in the LIMAH project and as considered in the search and hyperlinking task at MediaEval and TRECVid, are also prototypical links that can be made explicit for navigation. Pursuing work, we investigate two main issues:

- *Improving multimodal content-based linking.* We exploit achievements in entity linking to go beyond lexical or lexico-visual similarity and to provide semantic links that are easy to interpret for humans; carrying on, we work on link characterization, in search of mechanisms addressing link explainability (i.e., what is the nature of the link), for instance using attention models so as to focus

on the common parts of two documents or using natural language generation; a final topic that we address is that of linking textual content to external data sources in the field of journalism, e.g., leveraging topic models and cue phrases along with a short description of the external sources.

- *Dynamicity and user-adaptation.* One difficulty for explicit link creation is that links are often suited for one particular usage but not for another, thus requiring creating new links for each intended use; whereas link creation cannot be done online because of its computational cost, the alternative is to generate (almost) all possible links and provide users with selection mechanisms enabling personalization and user-adaptation in the exploration process; we design such strategies and investigate their impact on exploration tasks in search of a good trade-off between performance (few high-quality links) and genericity.

3.2.1.4.3. Summarizing.

Multimedia collections contain far too much information to allow any easy comprehension. It is mandatory to have facilities to aggregate and summarize a large body on information into a compact, concise and meaningful representation facilitating getting insight. Current technology suggests that multimedia content aggregation and story-telling are two complementary ways to provide users with such higher-level views. Yet, very few studies already investigated these issues. Recently, video or image captioning [87], [82] have been seen as a way to summarize visual content, opening the door to state-of-the-art multi-document text summarization [61] with text as a pivot modality. Automatic story-telling has been addressed for highly specific types of content, namely TV series [54] and news [73], [80], but still need a leap forward to be mostly automated, e.g., using constraint-based approaches for summarization [51], [80].

Furthermore, not only the original multimedia material has to be summarized, but the knowledge acquired from its analysis is also to summarize. It is important to be able to produce high-level views of the relationships between documents, emphasizing some structural distinguishing qualities. Graphs establishing such relationships need to be constructed at various level of granularity, providing some support for summarizing structural traits.

Summarizing multimedia information poses several scientific challenges that are:

- *Choosing the most relevant multimedia aggregation type*: Taking a multimedia collection into account, a same piece of information can be present in several modalities. The issue of selecting the most suitable one to express a given concept has thus to be considered together with the way to mix the various modalities into an acceptable production. Standard summarization algorithms have to be revisited so that they can handle continuous representation spaces, allowing them to benefit from the various modalities [55].
- *Expressing user's preferences*: Different users may appreciate quite different forms of multimedia summaries, and convenient ways to express their preferences have to be proposed. We for example focus on the opportunities offered by the constraint-based framework.
- *Evaluating multimedia summaries*: Finding criteria to characterize what a good summary is remains challenging, e.g., how to measure the global relevance of a multimodal summary and how to compare information between and across two modalities. We tackle this issue particularly via a collaboration with A. Smeaton at DCU, comparing the automatic measures we will develop to human judgments obtained by crowd-sourcing;
- *Taking into account structuring and dynamicity*: Typed links between multimedia fragments, and hierarchical topical structures of documents obtained via work previously developed within the team are two types of knowledge which have seldom been considered as long as summarization is concerned. Knowing that the event present in a document is causally related to another event described in another document can however modify the ways summarization algorithms have to consider information. Moreover the question of producing coarse-to-fine grain summaries exploiting the topical structure of documents is still an open issue. Summarizing dynamic collections is also challenging and it is one of the questions we consider.

LINKS Project-Team

3. Research Program

3.1. Background

The main objective of LINKS is to develop methods for querying and managing linked data collections. Even though open linked data is the most prominent example, we will focus on hybrid linked data collections, which are collections of semi-structured datasets in hybrid formats: graph-based, RDF, relational, and NoSQL. The elements of these datasets may be linked, either by pointers or by additional relations between the elements of the different datasets, for instance the "same-as" or "member-of" relations as in RDF.

The advantage of traditional data models is that there exist powerful querying methods and technologies that one might want to preserve. In particular, they come with powerful schemas that constraint the possible manners in which knowledge is represented to a finite number of patterns. The exhaustiveness of these patterns is essential for writing of queries that cover all possible cases. Pattern violations are excluded by schema validation. In contrast, RDF schema languages such as RDFS can only enrich the relations of a dataset by new relations, which also helps for query writing, but which cannot constraint the number of possible patterns, so that they do not come with any reasonable notion of schema validation.

The main weakness of traditional formats, however, is that they do not scale to large data collections as stored on the Web, while the RDF data models scales well to very big collections such as linked open data. Therefore, our objective is to study mixed data collections, some of which may be in RDF format, in which we can lift the advantages of smaller datasets in traditional formats to much larger linked data collections. Such data collections are typically distributed over the internet, where data sources may have rigid query facilities that cannot be easily adapted or extended.

The main assumption that we impose in order to enable the logical approach, is that the given linked data collection must be correct in most dimensions. This means that all datasets are well-formed with respect to their available constraints and schemas, and clean with respect to the data values in most of the components of the relations in the datasets. One of the challenges is to integrate good quality RDF datasets into this setting, another is to clean the incorrect data in those dimensions that are less proper. It remains to be investigated in how far these assumptions can be maintained in realistic applications, and how much they can be weakened otherwise.

For querying linked data collections, the main problems are to resolve the heterogeneity of data formats and schemas, to understand the efficiency and expressiveness of recursive queries, that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When linked data is dynamically created, exchanged, or updated, the problems are how to process linked data incrementally, and how to manage linked data collections that change dynamically. In any case (static and dynamic) one needs to find appropriate schema mappings for linking semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2. Querying Heterogeneous Linked Data

Our main objective is to query collections of linked datasets. In the static setting, we consider two kinds of links: explicit links between elements of the datasets, such as equalities or pointers, and logical links between relations of different datasets such as schema mappings. In the dynamic setting, we permit a third kind of links that point to "intentional" relations computable from a description, such as the application of a Web service or the application of a schema mapping.

We believe that collections of linked datasets are usually too big to ensure a global knowledge of all datasets. Therefore, schema mappings and constraints should remain between pairs of datasets. Our main goal is to be able to pose a query on a collection of datasets, while accounting for the possible recursive effects of schema mappings. For illustration, consider a ring of datasets D_1, D_2, D_3 linked by schema mappings M_1, M_2, M_3 that tell us how to complete a database D_i by new elements from the next database in the cycle.

The mappings M_i induce three intentional datasets I_1, I_2, and I_3, such that I_i contains all elements from D_i and all elements implied by M_i from the next intentional dataset in the ring:

$$I_1 = D_1 \cup M_1(I_2), \quad I_2 = D_2 \cup M_2(I_3), \quad I_3 = D_3 \cup M_3(I_1)$$

Clearly, the global information collected by the intentional datasets depends recursively on all three original datasets D_i. Queries to the global information can now be specified as standard queries to the intentional databases I_i. However, we will never materialize the intentional databases I_i. Instead, we can rewrite queries on one of the intentional datasets I_i to recursive queries on the union of the original datasets D_1, D_2, and D_3 with their links and relations. Therefore, a query answering algorithm is needed for recursive queries, that chases the "links" between the D_i in order to compute the part of I_i needed for the purpose of query answering.

This illustrates that we must account for the graph data models when dealing with linked data collections whose elements are linked, and that query languages for such graphs must provide recursion in order to chase links. Therefore, we will have to study graph databases with recursive queries, such as RDF graphs with SPARQL queries, but also other classes of graph databases and queries.

We study schemas and mappings between datasets with different kinds of data models and the complexity of evaluating recursive queries over graphs. In order to use schema mapping for efficiently querying the different datasets, we need to optimize the queries by taking into account the mappings. Therefore, we will study static analysis of schema mappings and recursive queries. Finally, we develop concrete applications in which our fundamental techniques can be applied.

3.3. Managing Dynamic Linked Data

With the quick growth of the information technology on the Web, more and more Web data gets created dynamically every day, for instance by smartphones, industrial machines, users of social networks, and all kinds of sensors. Therefore, large amounts of dynamic data need to be exchanged and managed by various data-centric web services, such as online shops, online newspapers, and social networks.

Dynamic data is often created by the application of some kind of service on the Web. This kind of data is intentional in the same spirit as the intentional data specified by the application of a schema mapping, or the application of some query to the hidden Web. Therefore, we will consider a third kind of links in the dynamic setting, that map to intentional data specified by whatever kind of function application. Such a function can be defined in data-centric programming languages, in the style of Active XML, XSLT, and NoSQL languages.

The dynamicity of data adds a further dimension to the challenges for linked data collections that we described before, while all the difficulties remain valid. One of the new aspects is that intentional data may be produced incrementally, as for instance when exchanged over data streams. Therefore, one needs incremental algorithms able to evaluate queries on incomplete linked data collections, that are extended or updated incrementally. Note that incremental data may be produced without end, such as a Twitter stream, so that one cannot wait for its completion. Instead, one needs to query and manage dynamic data with as low latency as possible. Furthermore, all static analysis problems are to be re-investigated in the presence of dynamic data.

Another aspect of dynamic data is distribution over the Web, and thus parallel processing as in the cloud. This raises the typical problems coming with data distribution: huge data sources cannot be moved without very high costs, while data must be replicated for providing efficient parallel access. This makes it difficult, if not impossible, to update replicated data consistently. Therefore, the consistency assumption has been removed by NoSQL databases for instance, while parallel algorithmic is limited to naive parallelization (i.e. map/reduce) where only few data needs to be exchanged.

We will investigate incremental query evaluation for distributed data-centered programming languages for linked data collections, dynamic updates as needed for linked data management, and static analysis for linked data workflows.

3.4. Linking Graphs

When datasets from independent sources are not linked with existing schema mappings, we would like to investigate symbolic machine learning solutions for inferring such mappings in order to define meaningful links between data from separate sources. This problem can be studied for various kinds of linked data collections. Before presenting the precise objectives, we will illustrate our approach on the example of linking data in two independent graphs: an address book of a research institute containing detailed personnel information and a (global) bibliographic database containing information on papers and their authors.

We remind that a schema allows to identify a collection of types each grouping objects from the same semantic class e.g., the collection of all persons in the address book and the collection of all authors in the bibliography database. As a schema is often lacking or underspecified in graph data models, we intend to investigate inference methods based on structural similarity of graph fragments used to describe objects from the same class in a given document e.g., in the bibliographic database every author has a name and a number of affiliations, while a paper has a title and a number of authors. Furthermore, our inference methods will attempt to identify, for every type, a set of possible keys, where by key we understand a collection of attributes of an object that uniquely identifies such an object in its semantic class. For instance, for a person in the address book two examples of a key are the name of the person and the office phone number of that person.

In the next step, we plan to investigate employing existing entity linkage solutions to identify pairs of types from different databases whose instances should be linked using compatible keys. For instance, persons in the address book should be linked with authors in the bibliographical database using the name as the compatible key. Linking the same objects (represented in different ways) in two databases can be viewed as an instance of a mapping between the two databases. Such mapping is, however, discriminatory because it typically maps objects from a specific subset of objects of given types. For instance, the mapping implied by linking persons in the address book with authors in the bibliographic database involves in fact researchers, a subgroup of personnel of the research institute, and authors affiliated with the research institute. Naturally, a subset of objects of a given type, or a subtype, can be viewed as a result of a query on the set of all objects, which on very basic level illustrates how learning data mappings can be reduced to learning queries.

While basic mappings link objects of the same type, more general mappings define how the same type of information is represented in two different databases. For instance, the email address and the postal address of an individual may be represented in one way in the address book and in another way in the bibliographic databases, and naturally, the query asking for the email address and the postal address of a person identified by a given name will differ from one database to the other. While queries used in the context of linking objects of compatible types are essentially unary, queries used in the context of linking information are n-ary and we plan to approach inference of general database mappings by investigating and employing algorithms for inference of n-ary queries.

An important goal in this research is elaborating a formal definition of *learnability* (feasibility of inference) of a given class of concepts (schemas of queries). We plan to following the example of Gold (1967), which requires not only the existence of an efficient algorithm that infers concepts consistent with the given input but the ability to infer every concept from the given class with a sufficiently informative input. Naturally, learnability depends on two parameters. The first parameter is the class of concepts i.e., a class of schema and

a class of queries, from which the goal concept is to be inferred. The second parameter is the type of input that an inference algorithm is given. This can be a set of examples of a concept e.g., instances of RDF databases for which we wish to construct a schema or a selection of nodes that a goal query is to select. Alternatively, a more general interactive scenario can be used where the learning algorithm inquires the user about the goal concept e.g., by asking to indicate whether a given node is to be selected or not (as membership queries of Angluin (1987)). In general, the richer the input is, the richer class of concepts can be handled, however, the richer class of queries is to be handled, the higher computational cost is to be expected. The primary task is to find a good compromise and identify classes of concepts that are of high practical value, allow efficient inference with possibly simple type of input.

The main open problem for graph-shaped data studied by Links are how to infer queries, schemas, and schema-mappings for graph-structured data.

LOKI Project-Team

3. Research Program

3.1. Introduction

Interaction is by nature a dynamic phenomenon that takes place between interactive systems and their users. Redesigning interactive systems to better account for interaction requires fine understanding of these dynamics from the user side so as to better handle them from the system side. In fact, layers of actual interactive systems abstract hardware and system resources from a system and programing perspective. Following our Interaction Machine concept, we are reconsidering these architectures from the user's perspective, through different *levels of dynamics of interaction* (see Figure 1).

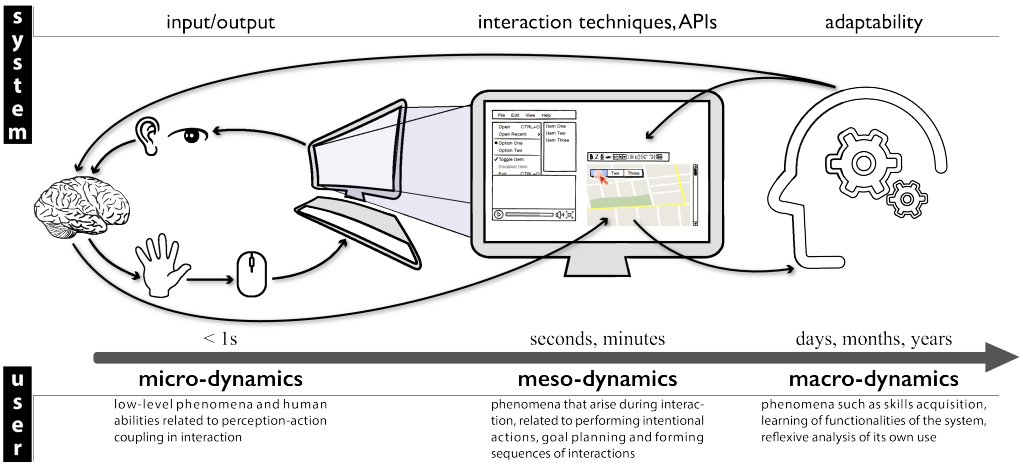


Figure 1. Levels of dynamics of interaction.

Considering phenomena that occur at each of these levels as well as their relationships will help us to acquire the necessary knowledge (Empowering Tools) and technological bricks (Interaction Machine) to reconcile the way interactive systems are designed and engineered with human abilities. Although our strategy is to investigate issues and address challenges for all of the three levels, our immediate priority is to focus on micro-dynamics since it concerns very fundamental knowledge about interaction and relates to very low-level parts of interactive systems, which is likely to influence our future research and developments at the other levels.

3.2. Micro-Dynamics

Micro-dynamics involve low-level phenomena and human abilities which are related to short time/instantness and to perception-action coupling in interaction, when the user has almost no control or consciousness of the action once it has been started. From a system perspective, it has implications mostly on input and output (I/O) management.

3.2.1. Transfer functions design and latency management

We have developed a recognized expertise in the characterization and the design of *transfer functions* [34], [45], i. e., the algorithmic transformations of raw user input for system use. Ideally, transfer functions should match the interaction context. Yet the question of how to maximize one or more criteria in a given context remains an open one, and on-demand adaptation is difficult because transfer functions are usually implemented at the lowest possible level to avoid latency. Latency has indeed long been known as a determinant of human performance in interactive systems [41] and recently regained attention with touch interactions [40]. These two problems require cross examination to improve performance with interactive systems: Latency can be a confounding factor when evaluating the effectiveness of transfer functions, and transfer functions can also include algorithms to compensate for latency.

We have recently proposed new cheap but robust methods for the measurement of end-to-end latency [2] and are currently working on compensation methods and the evaluation of their perceived side effects. Our goal is then to automatically adapt the transfer function to individual users and contexts of use while reducing latency in order to support stable and appropriate control. To achieve this, we will investigate combinations of low-level (embedded) and high-level (application) ways to take user capabilities and task characteristics into account and reduce or compensate for latency in different contexts, e. g., using a mouse or a touchpad, a touch-screen, an optical finger navigation device or a brain-computer interface. From an engineering perspective, this knowledge on low-level human factors will help us to rethink and redesign the I/O loop of interactive systems in order to better account for them and achieve more adapted and adaptable perception-action coupling.

3.2.2. Tactile feedback & haptic perception

We are also concerned with the physicality of human-computer interaction, with a focus on haptic perception and related technologies. For instance, when interacting with virtual objects such as software buttons on a touch surface, the user cannot feel the click sensation as with physical buttons. The tight coupling between how we perceive and how we manipulate objects is then essentially broken although this is instrumental for efficient direct manipulation. We have addressed this issue in multiple contexts by designing, implementing and evaluating novel applications of tactile feedback [5].

In comparison with many other modalities, one difficulty with tactile feedback is its diversity. It groups sensations of forces, vibrations, friction, or deformation. Although this is a richness, it also raises usability and technological challenges since each kind of haptic stimulation requires different kinds of actuators with their own parameters and thresholds. And results from one are hardly applicable to others. On a “knowledge” point of view, we want to better understand and empirically classify haptic variables and the kind of information they can represent (continuous, ordinal, nominal), their resolution, and their applicability to various contexts. From the “technology” perspective, we want to develop tools to inform and ease the design of haptic interactions taking best advantage of the different technologies in a consistent and transparent way.

3.3. Meso-Dynamics

Meso-dynamics relate to phenomena that arise during interaction, on a longer but still short time-scale. For users, it is related to performing intentional actions, to goal planning and tools selection, and to forming sequences of interactions based on a known set of rules or instructions. From the system perspective, it relates to how possible actions are exposed to the user and how they have to be executed (i. e., interaction techniques). It also has implication on the tools for designing and implementing those techniques (programming languages and APIs).

3.3.1. Interaction bandwidth and vocabulary

Interactive systems and their applications have an always-increasing number of available features and commands due to, e. g., the large amount of data to manipulate, increasing power and number of functionalities, or multiple contexts of use.

On the input side, we want to augment the *interaction bandwidth* between the user and the system in order to cope with this increasing complexity. In fact, most input devices capture only a few of the movements and actions the human body is capable of. Our arms and hands for instance have many degrees of freedom that are not fully exploited in common interfaces. We have recently designed new technologies to improve expressibility such as a bendable digitizer pen [36], or reliable technology for studying the benefits of finger identification on multi-touch interfaces [4].

On the output side, we want to expand users' *interaction vocabulary*. All of the features and commands of a system can not be displayed on screen at the same time and lots of *advanced* features are by default hidden to the users (e. g., hotkeys) or buried in deep hierarchies of command-triggering systems (e. g., menus). As a result, users tend to use only a subset of all the tools the system actually offers [44]. We will study how to help them to broaden their knowledge of available functions.

Through this "opportunistic" exploration of alternative and more expressive input methods and interaction techniques, we will particularly focus on the necessary technological requirements to integrate them into interactive systems, in relation with our redesign of the I/O stack at the micro-dynamics level.

3.3.2. Spatial and temporal continuity in interaction

At a higher-level, we will investigate how more expressive interaction techniques affect users' strategies when performing sequences of elementary actions and tasks. More generally, we will explore the "*continuity*" in interaction. Interactive systems have moved from one computer to multiple connected interactive devices (computer, tablets, phones, watches, etc.) that could also be augmented through a Mixed-Reality paradigm. This distribution of interaction raises new challenges from both usability and engineering perspectives that we clearly have to consider in our main objective of revisiting interactive systems [43]. It involves the simultaneous use of multiple devices and also the changes in the role of devices according to the location, the time, the task, and contexts of use: a tablet device can be used as the main device while traveling, and it becomes an input device or a secondary monitor for continuing the same task once in the office; a smart-watch can be used as a standalone device to send messages, but also as a remote controller for a wall-sized display. One challenge is then to design interaction techniques that support seamless and smooth transitions during these spatial and temporal changes of the system in order to maintain the continuity of uses and tasks, and how to integrate these principles in future interactive systems.

3.3.3. Expressive tools for prototyping, studying, and programming interaction

Current systems suffer from engineering issues that keep constraining and influencing how interaction is thought, designed, and implemented. Addressing the challenges we presented in this section and making the solutions possible require extended expressiveness, and researchers and designers must either wait for the proper toolkits to appear, or "hack" existing interaction frameworks, often bypassing existing mechanisms. For instance, numerous usability problems in existing interfaces stem from a common cause: the lack, or untimely discarding, of relevant information about how events are propagated and how changes come to occur in interactive environments. On top of our redesign of the I/O loop of interactive systems, we will investigate how to facilitate access to that information and also promote a more grounded and expressive way to describe and exploit input-to-output chains of events at every system level. We want to provide finer granularity and better-described connections between the *causes* of changes (e.g. input events and system triggers), their *context* (e.g. system and application states), their *consequences* (e.g. interface and data updates), and their *timing* [8]. More generally, a central theme of our Interaction Machine vision is to promote interaction as a first-class object of the system [33], and we will study alternative and better-adapted technologies for designing and programming interaction, such as we did recently to ease the prototyping of Digital Musical Instruments [1] or the programming of animations in graphical interfaces [10]. Ultimately, we want to propose a unified model of hardware and software scaffolding for interaction that will contribute to the design of our Interaction Machine.

3.4. Macro-Dynamics

Macro-dynamics involve longer-term phenomena such as skills acquisition, learning of functionalities of the system, reflexive analysis of its own use (e. g., when the user has to face novel or unexpected situations which require high-level of knowledge of the system and its functioning). From the system perspective, it implies to better support cross-application and cross-platform mechanisms so as to favor skill transfer. It also requires to improve the instrumentation and high-level logging capabilities to favor reflexive use, as well as flexibility and adaptability for users to be able to finely tune and shape their tools.

We want to move away from the usual binary distinction between "novices" and "experts" [3] and explore means to promote and assist digital skill acquisition in a more progressive fashion. Indeed, users have a permanent need to adapt their skills to the constant and rapid evolution of the tasks and activities they carry on a computer system, but also the changes in the software tools they use [47]. Software strikingly lacks powerful means of acquiring and developing these skills [3], forcing users to mostly rely on outside support (e. g., being guided by a knowledgeable person, following online tutorials of varying quality). As a result, users tend to rely on a surprisingly limited interaction vocabulary, or *make-do* with sub-optimal routines and tools [48]. Ultimately, the user should be able to master the interactive system to form durable and stabilized practices that would eventually become *automatic* and reduce the mental and physical efforts, making their interaction *transparent*.

In our previous work, we identified the fundamental factors influencing expertise development in graphical user interfaces, and created a conceptual framework that characterizes users' performance improvement with UIs [7], [3]. We designed and evaluated new command selection and learning methods to leverage user's digital skill development with user interfaces, on both desktop [6] and touch-based computers.

We are now interested in broader means to support the analytic use of computing tools:

- *to foster understanding of interactive systems.* As the digital world makes the shift to more and more complex systems driven by machine learning algorithms, we increasingly lose our comprehension of which process caused the system to respond in one way rather than another. We will study how novel interactive visualizations can help reveal and expose the "intelligence" behind, in ways that people better master their complexity.
- *to foster reflexion on interaction.* We will study how we can foster users' reflexion on their own interaction in order to encourage them to acquire novel digital skills. We will build real-time and off-line software for monitoring how user's ongoing activity is conducted at an application and system level. We will develop augmented feedbacks and interactive history visualization tools that will offer contextual visualizations to help users to better understand and share their activity, compare their actions to that of others, and discover possible improvement.
- *to optimize skill-transfer and tool re-appropriation.* The rapid evolution of new technologies has drastically increased the frequency at which systems are updated, often requiring to relearn everything from scratch. We will explore how we can minimize the cost of having to appropriate an interactive tool by helping users to capitalize on their existing skills.

We plan to explore these questions as well as the use of such aids in several contexts like web-based, mobile, or BCI-based applications. Although, a core aspect of this work will be to design systems and interaction techniques that will be as little platform-specific as possible, in order to better support skill transfer. Following our Interaction Machine vision, this will lead us to rethink how interactive systems have to be engineered so that they can offer better instrumentation, higher adaptability, and fewer separation between applications and tasks in order to support reuse and skill transfer.

MAGNET Project-Team

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data in order to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. We consider information networks in which the data consist of both feature vectors and texts. We model such networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new online and batch learning algorithms for various problems (node classification / clustering, link classification / prediction) which exploit the relationships between data entities and, overall, the graph topology. We are also interested in searching for the best hidden graph structure to be generated for solving a given learning task. Our research will be based on generative models for graphs, on machine learning for graphs and on machine learning for texts. The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling, randomization and active learning can be leveraged to improve the scalability of the proposed algorithms.

Our research program is organized according to the following questions:

1. How to go beyond vectorial classification models in Natural Language Processing (NLP) tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create networks from observations of information diffusion processes?
3. How to design methods able to achieve a good trade-off between predictive accuracy and computational complexity?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

3.2. Beyond Vectorial Models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Interestingly, NLP research has been rather slow in coming to terms with this situation, and most of the literature still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [38], [41].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NLP tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative to, or at least complement, structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [21], [43].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performance for several NLP tasks. We think that a "network effect", similar to the one that took place in Information Retrieval (with the PageRank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [42].

Part of the challenge here will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, friendship, or thread membership. Unfortunately, modeling NLP problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [21], [46]. We identify the issue of adaptive graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3 .

As noted above, many NLP tasks have been recast as structured prediction problems, allowing to capture (some of the) output dependencies. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph regularization and graph propagation methods. Within such approaches, labels are typically binary or in a small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [43], [31]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specified tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NLP problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [33].

The NLP tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisioned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that they are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team (see for instance [32], [28], [30]). As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such as sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [45].

We have already initiated some work on the coreference resolution with graph-based learning, by casting the problem as an instance of spectral clustering [30].

3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data modeling process and convey crucially important information for classifying nodes, making it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to several classification problems is to represent the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data [40], face recognition [29], and text categorization [34].

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example for the TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the χ^2 distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy ([47], [22], [23]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. The first one is about how to choose the best similarity measure given the objective learning task. This question is related to the question of metric and similarity learning ([24], [25]) which has not been considered in the context of graph-based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-k outcomes of greedy algorithms.

A different way we envision adaptive graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [36]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provide a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recovered and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs

where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labeling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2 , but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find a smooth labeling function corresponding to an harmonic function on both manifolds in input and output.

Scalability is one of the main issues in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computational complexity scales quadratically, or slower, in the number of considered data objects (usually nodes or edges, depending on the task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting. A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [44].

At the level of mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [35], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [26]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

Another way to tackle scalability problems is to exploit the inherent decentralized nature of very large graphs. Indeed, in many situations very large graphs are the abstract view of the digital activities of a very large set of users equipped with their own device. Nowadays, smartphones, tablets and even sensors have storage and computation power and gather a lot of data that serve to analytics, prediction, suggestion and personalized recommendation. Gathering all user data in large data centers is costly because it requires oversized infrastructures with huge energy consumption and large bandwidth networks. Even though cloud architectures can optimize such infrastructures, data concentration is also prone to security leaks, lost of privacy and data governance for end users. The alternative we have started to develop in Magnet is to devise decentralized, private and personalized machine learning algorithms so that they can be deployed in the personal devices. The key challenges are therefore to learn in a collaborative way in a network of learners and to preserve privacy and control on personal data.

3.5. Beyond Homophilic Relationships

In many cases, algorithms for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ([27], [37]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing common interests is a key reason for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical model, called signed graphs, has an unexpectedly rich additional complexity. For example, their spectral properties, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of graphs. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting signs of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationships between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [39]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting scheme allows to weigh the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This condition provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes (the theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

MAGRIT Team

3. Research Program

3.1. Matching and 3D tracking

One of the most basic problems currently limiting AR applications is the registration problem. The objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.

As a large number of potential AR applications are interactive, real time pose computation is required. Although the registration problem has received a lot of attention in the computer vision community, the problem of real-time registration is still far from being a solved problem, especially for unstructured environments. Ideally, an AR system should work in all environments, without the need to prepare the scene ahead of time, independently of the variations in experimental conditions (lighting, weather condition,...)

For several years, the MAGRIT project has been aiming at developing on-line and marker-less methods for camera pose computation. The main difficulty with on-line tracking is to ensure robustness of the process over time. For off-line processes, robustness is achieved by using spatial and temporal coherence of the considered sequence through move-matching techniques. To get robust open-loop systems, we have investigated various methods, ranging from statistical methods to the use of hybrid camera/sensor systems. Many of these methods are dedicated to piecewise-planar scenes and combine the advantage of move-matching methods and model-based methods. In order to reduce statistical fluctuations in viewpoint computation, which lead to unpleasant jittering or sliding effects, we have also developed model selection techniques which allow us to noticeably improve the visual impression and to reduce drift over time. Another line of research which has been considered in the team to improve the reliability and the robustness of pose algorithms is to combine the camera with another form of sensor in order to compensate for the shortcomings of each technology.

The success of pose computation over time largely depends on the quality of the matching at the initialization stage. Indeed, the current image may be very different from the appearances described in the model both on the geometrical and the photometric sides. Research is thus conducted in the team on the use of probabilistic methods to establish robust correspondences of features. The use of *a contrario* methods has been investigated to achieve this aim [7]. We especially addressed the complex case of matching in scenes with repeated patterns which are common in urban scenes. We are also investigating the problem of matching images taken from very different viewpoints which is central for the re-localization issue in AR. Within the context of a scene model acquired with structure-from-motion techniques, we are currently investigating the use of viewpoint simulation in order to allow successful pose computation even if the considered image is far from the positions used to build the model [15].

Recently, the issue of tracking deformable objects has gained importance in the team. This topic is mainly addressed in the context of medical applications through the design of bio-mechanical models guided by visual features [2]. We have successfully investigated the use of such models in laparoscopy, with a vascularized model of the liver and with a hyper-elastic model for tongue tracking in ultrasound images. However, these results have been obtained so far in relatively controlled environments, with non-pathological cases. When clinical routine applications are to be considered, many parameters and considerations need to be taken into account. Among the problems that need to be addressed are more realistic model representations, the specification of the range of physical parameters and the need to enforce the robustness of the tracking with respect to outliers, which are common in the interventional context.

3.2. Image-based Modeling

Modeling the scene is a fundamental issue in AR for many reasons. First, pose computation algorithms often use a model of the scene or at least some 3D knowledge on the scene. Second, effective AR systems require a

model of the scene to support interactions between the virtual and the real objects such as occlusions, lighting reflections, contacts... in real-time. Unlike pose computation which has to be performed in a sequential way, scene modeling can be considered as an off-line or an on-line problem depending on the requirements of the targeted application. Interactive in-situ modeling techniques have thus been developed with the aim to enable the user to define what is relevant at the time the model is being built during the application. On the other hand, we also proposed off-line multimodal techniques, mainly dedicated to AR medical applications, with the aim of obtaining realistic and possibly dynamic models of organs suitable for real-time simulation [3].

In-situ modeling
In-situ modeling allows a user to directly build a 3D model of his/her surrounding environment and verify the geometry against the physical world in real-time. This is of particular interest when using AR in unprepared environments or building scenes that either have an ephemeral existence (e.g., a film set) or cannot be accessed frequently (e.g., a nuclear power plant). We have especially investigated two systems, one based on the image content only and the other based on multiple data coming from different sensors (camera, inertial measurement unit, laser rangefinder). Both systems use the camera-mouse principle [34] (i.e., interactions are performed by aiming at the scene through a video camera) and both systems have been designed to acquire polygonal textured models, which are particularly useful for camera tracking and object insertion in AR.

Multimodal modeling for real-time simulation
With respect to classical AR applications, AR in medical context differs in the nature and the size of the data which are available: a large amount of multimodal data is acquired on the patient or possibly on the operating room through sensing technologies or various image acquisitions [32]. The challenge is to analyze these data, to extract interesting features, to fuse and to visualize this information in a proper way. Within the MAGRIT team, we address several key problems related to medical augmented environments. Being able to acquire multimodal data which are temporally synchronized and spatially registered is the first difficulty we face when considering medical AR. Another key requirement of AR medical systems is the availability of 3D (+t) models of the organ/patient built from images, to be overlaid onto the users' view of the environment.

Methods for multimodal modeling are strongly dependent on the imaging modalities and the organ specificities. We thus only address a restricted number of medical applications –interventional neuro-radiology, laparoscopic surgery– for which we have a strong expertise and close relationships with motivated clinicians. In these applications, our aim is to produce realistic models and then realistic simulations of the patient to be used for the training of surgeons or the re-education of patients.

One of our main applications is about neuroradiology. For the last 20 years, we have been working in close collaboration with the neuroradiology laboratory (CHRU-University Hospital of Nancy) and GE Healthcare. As several imaging modalities are now available in an intraoperative context (2D and 3D angiography, MRI, ...), our aim is to develop a multi-modality framework to assist therapeutic decision and treatment.

We have mainly been interested in the effective use of a multimodality framework in the treatment of arteriovenous malformations (AVM) and aneurysms in the context of interventional neuroradiology. The goal of interventional gestures is to guide endoscopic tools towards the pathology with the aim to perform embolization of the AVM or to fill the aneurysmal cavity by placing coils. We have proposed and developed multimodality and augmented reality tools which make various image modalities (2D and 3D angiography, fluoroscopic images, MRI, ...) cooperate in order to assist physicians in clinical routine. One of the successes of this collaboration is the implementation of the concept of *augmented fluoroscopy*, which helps the surgeon to guide endoscopic tools towards the pathology. Lately, in cooperation with the team MIMESIS, we have proposed new methods for implicit modeling of the vasculature with the aim of obtaining near real-time simulation of the coil deployment in the aneurysm [3]. These works open the way towards near real-time patient-based simulations of interventional gestures both for training and for planning.

3.3. Parameter estimation

Many problems in computer vision or image analysis can be formulated in terms of parameter estimation from image-based measurements. This is the case of many problems addressed in the team such as pose

computation or image-guided estimation of 3D deformable models. Often traditional robust techniques which take into account the covariance on the measurements are sufficient to achieve reliable parameter estimation. However, depending on their number, their spatial distribution and the uncertainty on these measurements, some problems are very sensitive to noise and there is a considerable interest in considering how parameter estimation could be improved if additional information on the noise were available. Another common problem in our field of research is the need to estimate constitutive parameters of the models, such as (bio)-mechanical parameters for instance. Direct measurement methods are destructive, and elaborating image-based methods is thus highly desirable. Besides designing appropriate estimation algorithms, a fundamental question is to understand what group of parameters under study can be reliably estimated from a given experimental setup.

This line of research is relatively new in the team. One of the challenges is to improve image-based parameter estimation techniques considering sensor noise and specific image formation models. In a collaboration with the Pascal Institute (Clermont Ferrand), metrological performance enhancement for experimental solid mechanics has been addressed through the development of dedicated signal processing methods [6]. In the medical field, specific methods based on an adaptive evolutionary optimization strategy have been designed for estimating respiratory parameters [8]. In the context of designing realistic simulators for neuroradiology, we are now considering how parameters involved in the simulation could be adapted to fit real images.

MANAO Project-Team

3. Research Program

3.1. Related Scientific Domains

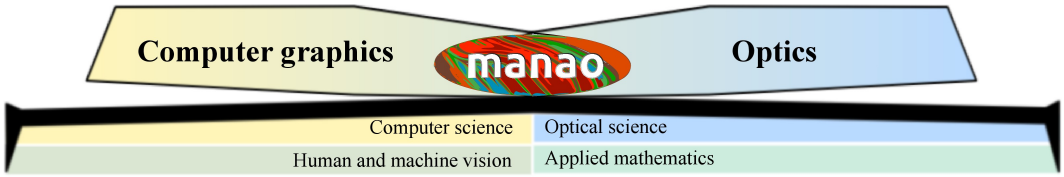


Figure 3. Related scientific domains of the MANAO project.

The *MANAO* project aims at studying, acquiring, modeling, and rendering the interactions between the three components that are light, shape, and matter from the viewpoint of an observer. As detailed more lengthily in the next section, such a work will be done using the following approach: first, we will tend to consider that these three components do not have strict frontiers when considering their impacts on the final observers; then, we will not only work in **computer graphics**, but also at the intersection of computer graphics and **optics**, exploring the mutual benefits that the two domains may provide. It is thus intrinsically a **transdisciplinary** project (as illustrated in Figure 3) and we expect results in both domains.

Thus, the proposed team-project aims at establishing a close collaboration between computer graphics (e.g., 3D modeling, geometry processing, shading techniques, vector graphics, and GPU programming) and optics (e.g., design of optical instruments, and theories of light propagation). The following examples illustrate the strengths of such a partnership. First, in addition to simpler radiative transfer equations [36] commonly used in computer graphics, research in the later will be based on state-of-the-art understanding of light propagation and scattering in real environments. Furthermore, research will rely on appropriate instrumentation expertise for the measurement [48], [49] and display [47] of the different phenomena. Reciprocally, optics researches may benefit from the expertise of computer graphics scientists on efficient processing to investigate interactive simulation, visualization, and design. Furthermore, new systems may be developed by unifying optical and digital processing capabilities. Currently, the scientific background of most of the team members is related to computer graphics and computer vision. A large part of their work have been focused on simulating and analyzing optical phenomena as well as in acquiring and visualizing them. Combined with the close collaboration with the optics laboratory LP2N (http://www.lp2n.fr) and with the students issued from the "Institut d'Optique" (http://www.institutoptique.fr), this background ensures that we can expect the following results from the project: the construction of a common vocabulary for tightening the collaboration between the two scientific domains and creating new research topics. By creating this context, we expect to attract (and even train) more trans-disciplinary researchers.

At the boundaries of the *MANAO* project lie issues in **human and machine vision**. We have to deal with the former whenever a human observer is taken into account. On one side, computational models of human vision are likely to guide the design of our algorithms. On the other side, the study of interactions between light, shape, and matter may shed some light on the understanding of visual perception. The same kind of connections are expected with machine vision. On the one hand, traditional computational methods for acquisition (such as photogrammetry) are going to be part of our toolbox. On the other hand, new display technologies (such as the ones used for augmented reality) are likely to benefit from our integrated approach

and systems. In the *MANAO* project we are mostly users of results from human vision. When required, some experimentation might be done in collaboration with experts from this domain, like with the European PRISM project. For machine vision, provided the tight collaboration between optical and digital systems, research will be carried out inside the *MANAO* project.

Analysis and modeling rely on **tools from applied mathematics** such as differential and projective geometry, multi-scale models, frequency analysis [38] or differential analysis [70], linear and non-linear approximation techniques, stochastic and deterministic integrations, and linear algebra. We not only rely on classical tools, but also investigate and adapt recent techniques (e.g., improvements in approximation techniques), focusing on their ability to run on modern hardware: the development of our own tools (such as Eigen) is essential to control their performances and their abilities to be integrated into real-time solutions or into new instruments.

3.2. Research axes

The *MANAO* project is organized around four research axes that cover the large range of expertise of its members and associated members. We briefly introduce these four axes in this section. More details and their inter-influences that are illustrated in the Figure 2 will be given in the following sections.

Axis 1 is the theoretical foundation of the project. Its main goal is to increase the understanding of light, shape, and matter interactions by combining expertise from different domains: optics and human/machine vision for the analysis and computer graphics for the simulation aspect. The goal of our analyses is to identify the different layers/phenomena that compose the observed signal. In a second step, the development of physical simulations and numerical models of these identified phenomena is a way to validate the pertinence of the proposed decompositions.

In Axis 2, the final observers are mainly physical captors. Our goal is thus the development of new acquisition and display technologies that combine optical and digital processes in order to reach fast transfers between real and digital worlds, in order to increase the convergence of these two worlds.

Axes 3 and 4 focus on two aspects of computer graphics: rendering, visualization and illustration in Axis 3, and editing and modeling (content creation) in Axis 4. In these two axes, the final observers are mainly human users, either generic users or expert ones (e.g., archaeologist [74], computer graphics artists).

3.3. Axis 1: Analysis and Simulation

Challenge: Definition and understanding of phenomena resulting from interactions between light, shape, and matter as seen from an observer point of view.

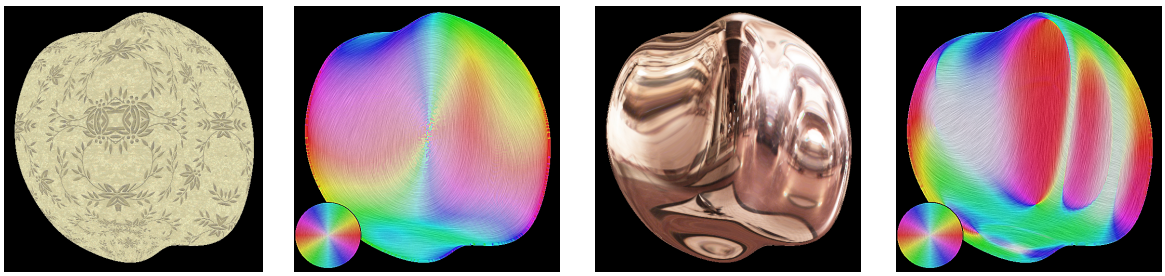
Results: Theoretical tools and numerical models for analyzing and simulating the observed optical phenomena.

To reach the goals of the *MANAO* project, we need to **increase our understanding** of how light, shape, and matter act together in synergy and how the resulting signal is finally observed. For this purpose, we need to identify the different phenomena that may be captured by the targeted observers. This is the main objective of this research axis, and it is achieved by using three approaches: the simulation of interactions between light, shape, and matter, their analysis and the development of new numerical models. This resulting improved knowledge is a foundation for the researches done in the three other axes, and the simulation tools together with the numerical models serve the development of the joint optical/digital systems in Axis 2 and their validation.

One of the main and earliest goals in computer graphics is to faithfully reproduce the real world, focusing mainly on light transport. Compared to researchers in physics, researchers in computer graphics rely on a subset of physical laws (mostly radiative transfer and geometric optics), and their main concern is to efficiently use the limited available computational resources while developing as fast as possible algorithms. For this purpose, a large set of theoretical as well as computational tools has been introduced to take a **maximum benefit of hardware** specificities. These tools are often dedicated to specific phenomena (e.g., direct or indirect lighting, color bleeding, shadows, caustics). An efficiency-driven approach needs such a classification of light paths [44] in order to develop tailored strategies [86]. For instance, starting from simple direct lighting,

more complex phenomena have been progressively introduced: first diffuse indirect illumination [42], [78], then more generic inter-reflections [51], [36] and volumetric scattering [75], [33]. Thanks to this search for efficiency and this classification, researchers in computer graphics have developed a now recognized expertise in fast-simulation of light propagation. Based on finite elements (radiosity techniques) or on unbiased Monte Carlo integration schemes (ray-tracing, particle-tracing, ...), the resulting algorithms and their combination are now sufficiently accurate to be used-back in physical simulations. The *MANAO* project will continue the search for **efficient and accurate simulation** techniques, but extending it from computer graphics to optics. Thanks to the close collaboration with scientific researchers from optics, new phenomena beyond radiative transfer and geometric optics will be explored.

Search for algorithmic efficiency and accuracy has to be done in parallel with **numerical models**. The goal of visual fidelity (generalized to accuracy from an observer point of view in the project) combined with the goal of efficiency leads to the development of alternative representations. For instance, common classical finite-element techniques compute only basis coefficients for each discretization element: the required discretization density would be too large and to computationally expensive to obtain detailed spatial variations and thus visual fidelity. Examples includes texture for decorrelating surface details from surface geometry and high-order wavelets for a multi-scale representation of lighting [32]. The numerical complexity explodes when considering directional properties of light transport such as radiance intensity (Watt per square meter and per steradian - $W.m^{-2}.sr^{-1}$), reducing the possibility to simulate or accurately represent some optical phenomena. For instance, Haar wavelets have been extended to the spherical domain [77] but are difficult to extend to non-piecewise-constant data [80]. More recently, researches prefer the use of Spherical Radial Basis Functions [83] or Spherical Harmonics [69]. For more complex data, such as reflective properties (e.g., BRDF [63], [52] - 4D), ray-space (e.g., Light-Field [60] - 4D), spatially varying reflective properties (6D - [73]), new models, and representations are still investigated such as rational functions [66] or dedicated models [20] and parameterizations [76], [81]. For each (newly) defined phenomena, we thus explore the space of possible numerical representations to determine the **most suited one for a given application**, like we have done for BRDF [66].



Texuring | 1st order gradient field | Environment reflection | 2st order gradient field

Figure 4. First-oder analysis [87] have shown that shading variations are caused by depth variations (first-order gradient field) and by normal variations (second-order fields). These fields are visualized using hue and saturation to indicate direction and magnitude of the flow respectively.

Before being able to simulate or to represent the different **observed phenomena**, we need to define and describe them. To understand the difference between an observed phenomenon and the classical light, shape, and matter decomposition, we can take the example of a highlight. Its observed shape (by a human user or a sensor) is the resulting process of the interaction of these three components, and can be simulated this way. However, this does not provide any intuitive understanding of their relative influence on the final shape: an artist will directly describe the resulting shape, and not each of the three properties. We thus want to decompose the observed signal into models for each scale that can be easily understandable, representable,

and manipulable. For this purpose, we will rely on the **analysis** of the resulting interaction of light, shape, and matter as observed by a human or a physical sensor. We first consider this analysis from an **optical point of view**, trying to identify the different phenomena and their scale according to their mathematical properties (e.g., differential [70] and frequency analysis [38]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 4). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [39] both to evaluate the results and to guide the simulations.

3.4. Axis 2: From Acquisition to Display

Challenge: Convergence of optical and digital systems to blend real and virtual worlds.

Results: Instruments to acquire real world, to display virtual world, and to make both of them interact.

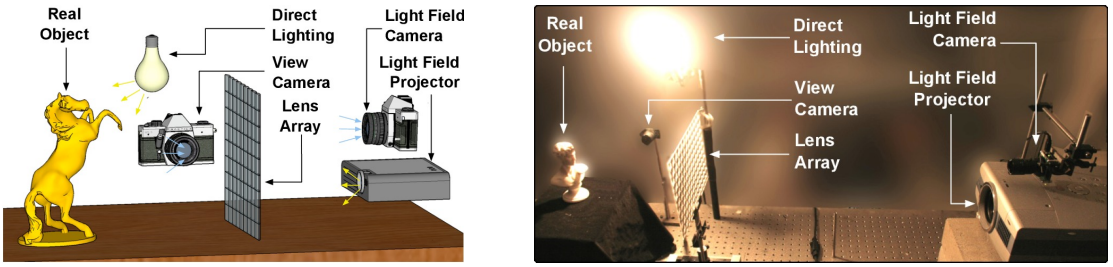


Figure 5. Light-Field transfer: global illumination between real and synthetic objects [31]

In this axis, we investigate *unified acquisition and display systems*, that is systems which combine optical instruments with digital processing. From digital to real, we investigate new display approaches [60], [47]. We consider projecting systems and surfaces [27], for personal use, virtual reality and augmented reality [22]. From the real world to the digital world, we favor direct measurements of parameters for models and representations, using (new) optical systems unless digitization is required [41], [40]. These resulting systems have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [45], [21], [46], [49]. By efficient, we mean that we want to shorten the path between the real world and the virtual world by increasing the data bandwidth between the real (analog) and the virtual (digital) worlds, and by reducing the latency for real-time interactions (we have to prevent unnecessary conversions, and to reduce processing time). To reach this goal, the systems have to be designed as a whole, not by a simple concatenation of optical systems and digital processes, nor by considering each component independently [50].

To increase data bandwidth, one solution is to **parallelize more and more the physical systems**. One possible solution is to multiply the number of simultaneous acquisitions (e.g., simultaneous images from multiple viewpoints [49], [68]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [60]. However, full acquisition or display of 3D real environments theoretically requires a continuous field of viewpoints, leading to huge data size. Despite the current belief that the increase of computational power will fill the missing gap, when it comes to visual or physical realism, if you double the processing power, people may want four times more accuracy, thus increasing data size as well. To reach the best performances, a trade-off has to be found between the amount of data required to represent accurately the reality and the amount of required processing. This trade-off may be achieved using **compressive sensing**. Compressive sensing is a new trend issued from the applied mathematics community that provides tools to accurately reconstruct a signal from a small set of measurements assuming that it is sparse in a transform domain (e.g., [67], [92]).

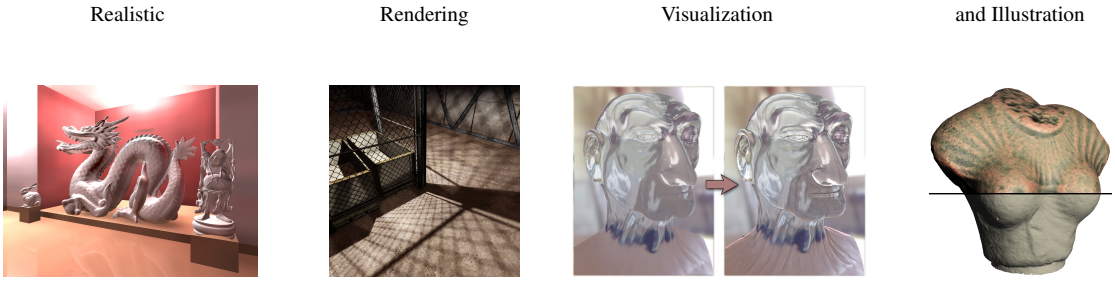
We prefer to achieve this goal by avoiding as much as possible the classical approach where acquisition is followed by a fitting step: this requires in general a large amount of measurements and the fitting itself may consume consequently too much memory and preprocessing time. By **preventing unnecessary conversion** through fitting techniques, such an approach increase the speed and reduce the data transfer for acquisition but also for display. One of the best recent examples is the work of Cossairt et al. [31]. The whole system is designed around a unique representation of the energy-field issued from (or leaving) a 3D object, either virtual or real: the Light-Field. A Light-Field encodes the light emitted in any direction from any position on an object. It is acquired thanks to a lens-array that leads to the capture of, and projection from, multiple simultaneous viewpoints. A unique representation is used for all the steps of this system. Lens-arrays, parallax barriers, and coded-aperture [57] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [0] [50] and acquisition of light-sources [41]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [55]. More generally, by designing unified optical and digital systems [64], it is possible to leverage the requirement of processing power, the memory footprint, and the cost of optical instruments.

Those are only some examples of what we investigate. We also consider the following approaches to develop new unified systems. First, similar to (and based on) the analysis goal of Axis 1, we have to take into account as much as possible the characteristics of the measurement setup. For instance, when fitting cannot be avoided, integrating them may improve both the processing efficiency and accuracy [66]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [58]). Finally, the experience of the group in surface modeling help the design of optical surfaces [53] for light sources or head-mounted displays.

3.5. Axis 3: Rendering, Visualization and Illustration

Challenge: How to offer the most legible signal to the final observer in real-time?

Results: High-level shading primitives, expressive rendering techniques for object depiction, real-time realistic rendering algorithms



(a) Global illumination [65] (b) Shadows [93] (c) Shape enhancement [89] (d) Shape depiction [19]

Figure 6. In the MANAO project, we are investigating rendering techniques from realistic solutions (e.g., inter-reflections (a) and shadows (b)) to more expressive ones (shape enhancement (c) with realistic style and shape depiction (d) with stylized style) for visualization.

The main goal of this axis is to offer to the final observer, in this case mostly a human user, the most legible signal in real-time. Thanks to the analysis and to the decomposition in different phenomena resulting from interactions between light, shape, and matter (Axis 1), and their perception, we can use them to convey essential information in the most pertinent way. Here, the word *pertinent* can take various forms depending on the application.

[0] Lytro, http://www.lytro.com/

In the context of scientific illustration and visualization, we are primarily interested in tools to convey shape or material characteristics of objects in animated 3D scenes. **Expressive rendering** techniques (see Figure 6 c,d) provide means for users to depict such features with their own style. To introduce our approach, we detail it from a shape-depiction point of view, domain where we have acquired a recognized expertise. Prior work in this area mostly focused on stylization primitives to achieve line-based rendering [90], [54] or stylized shading [25], [89] with various levels of abstraction. A clear representation of important 3D **object features** remains a major challenge for better shape depiction, stylization and abstraction purposes. Most existing representations provide only local properties (e.g., curvature), and thus lack characterization of broader shape features. To overcome this limitation, we are developing higher level descriptions of shape [18] with increased robustness to sparsity, noise, and outliers. This is achieved in close collaboration with Axis 1 by the use of higher-order local fitting methods, multi-scale analysis, and global regularization techniques. In order not to neglect the observer and the material characteristics of the objects, we couple this approach with an analysis of the appearance model. To our knowledge, this is an approach which has not been considered yet. This research direction is at the heart of the *MANAO* project, and has a strong connection with the analysis we plan to conduct in Axis 1. Material characteristics are always considered at the light ray level, but an understanding of **higher-level primitives** (like the shape of highlights and their motion) would help us to produce more legible renderings and permit novel stylizations; for instance, there is no method that is today able to create stylized renderings that follow the motion of highlights or shadows. We also believe such tools also play a fundamental role for geometry processing purposes (such as shape matching, reassembly, simplification), as well as for editing purposes as discussed in Axis 4.

In the context of **real-time photo-realistic rendering** ((see Figure 6 a,b), the challenge is to compute the most plausible images with minimal effort. During the last decade, a lot of work has been devoted to design approximate but real-time rendering algorithms of complex lighting phenomena such as soft-shadows [91], motion blur [38], depth of field [79], reflexions, refractions, and inter-reflexions. For most of these effects it becomes harder to discover fundamentally new and faster methods. On the other hand, we believe that significant speedup can still be achieved through more clever use of **massively parallel architectures** of the current and upcoming hardware, and/or through more clever tuning of the current algorithms. In particular, regarding the second aspect, we remark that most of the proposed algorithms depend on several parameters which can be used to **trade the speed over the quality**. Significant speed-up could thus be achieved by identifying effects that would be masked or facilitated and thus devote appropriate computational resources to the rendering [56], [37]. Indeed, the algorithm parameters controlling the quality vs speed are numerous without a direct mapping between their values and their effect. Moreover, their ideal values vary over space and time, and to be effective such an auto-tuning mechanism has to be extremely fast such that its cost is largely compensated by its gain. We believe that our various work on the analysis of the appearance such as in Axis 1 could be beneficial for such purpose too.

Realistic and real-time rendering is closely related to Axis 2: real-time rendering is a requirement to close the loop between real world and digital world. We have to thus develop algorithms and rendering primitives that allow the integration of the acquired data into real-time techniques. We have also to take care of that these real-time techniques have to work with new display systems. For instance, stereo, and more generally multi-view displays are based on the multiplication of simultaneous images. Brute force solutions consist in independent rendering pipeline for each viewpoint. A more energy-efficient solution would take advantages of the computation parts that may be factorized. Another example is the rendering techniques based on image processing, such as our work on augmented reality [29]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [47] would require new rendering pipelines.

3.6. Axis 4: Editing and Modeling

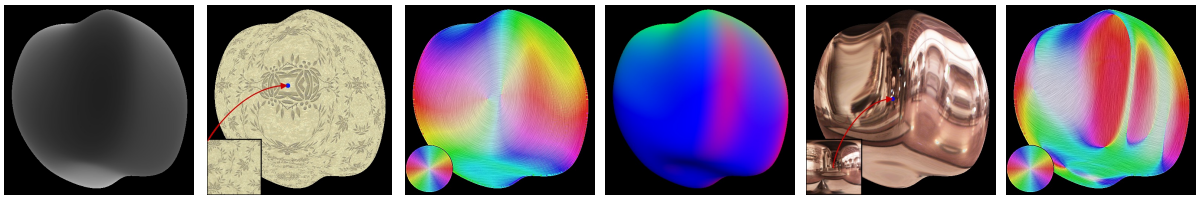
Challenge: Editing and modeling appearance using drawing- or sculpting-like tools through high level representations.

Results: High-level primitives and hybrid representations for appearance and shape.

During the last decade, the domain of computer graphics has exhibited tremendous improvements in image quality, both for 2D applications and 3D engines. This is mainly due to the availability of an ever increasing amount of shape details, and sophisticated appearance effects including complex lighting environments. Unfortunately, with such a growth in visual richness, even so-called *vectorial* representations (e.g., subdivision surfaces, Bézier curves, gradient meshes, etc.) become very dense and unmanageable for the end user who has to deal with a huge mass of control points, color labels, and other parameters. This is becoming a major challenge, with a necessity for novel representations. This Axis is thus complementary of Axis 3: the focus is the development of primitives that are easy to use for modeling and editing.

More specifically, we plan to investigate *vectorial representations* that would be amenable to the production of rich shapes with a minimal set of primitives and/or parameters. To this end we plan to build upon our insights on dynamic local reconstruction techniques and implicit surfaces [30] [24]. When working in 3D, an interesting approach to produce detailed shapes is by means of procedural geometry generation. For instance, many natural phenomena like waves or clouds may be modeled using a combination of procedural functions. Turning such functions into triangle meshes (main rendering primitives of GPUs) is a tedious process that appears not to be necessary with an adapted vectorial shape representation where one could directly turn procedural functions into implicit geometric primitives. Since we want to prevent unnecessary conversions in the whole pipeline (here, between modeling and rendering steps), we will also consider *hybrid representations* mixing meshes and implicit representations. Such research has thus to be conducted while considering the associated editing tools as well as performance issues. It is indeed important to keep *real-time performance* (cf. Axis 2) throughout the interaction loop, from user inputs to display, via editing and rendering operations. Finally, it would be interesting to add *semantic information* into 2D or 3D geometric representations. Semantic geometry appears to be particularly useful for many applications such as the design of more efficient manipulation and animation tools, for automatic simplification and abstraction, or even for automatic indexing and searching. This constitutes a complementary but longer term research direction.

In the *MANAO* project, we want to investigate representations beyond the classical light, shape, and matter decomposition. We thus want to directly control the appearance of objects both in 2D and 3D applications (e.g., [84]): this is a core topic of computer graphics. When working with 2D vector graphics, digital artists must carefully set up color gradients and textures: examples range from the creation of 2D logos to the photo-realistic imitation of object materials. Classic vector primitives quickly become impractical for creating illusions of complex materials and illuminations, and as a result an increasing amount of time and skill is required. This is only for still images. For animations, vector graphics are only used to create legible appearances composed of simple lines and color gradients. There is thus a need for more complex primitives that are able to accommodate complex reflection or texture patterns, while keeping the ease of use of vector graphics. For instance, instead of drawing color gradients directly, it is more advantageous to draw flow lines that represent local surface concavities and convexities. Going through such an intermediate structure then allows to deform simple material gradients and textures in a coherent way (see Figure 7), and animate them all at once. The manipulation of 3D object materials also raises important issues. Most existing material models are tailored to faithfully reproduce physical behaviors, not to be *easily controllable* by artists. Therefore artists learn to tweak model parameters to satisfy the needs of a particular shading appearance, which can quickly become cumbersome as the complexity of a 3D scene increases. We believe that an alternative approach is required, whereby material appearance of an object in a typical lighting environment is directly input (e.g., painted or drawn), and adapted to match a plausible material behavior. This way, artists will be able to create their own appearance (e.g., by using our shading primitives [84]), and replicate it to novel illumination environments and 3D models. For this purpose, we will rely on the decompositions and tools issued from Axis 1.



(a) (b) (c) (d) (e) (f)

Figure 7. Based on our analysis [87] (Axis 1), we have designed a system that mimics texture (left) and shading (right) effects using image processing alone. It takes depth (a) and normal (d) images as input, and uses them to deform images (b-e) in ways that closely approximate surface flows (c-f). It provides a convincing, yet artistically controllable illusion of 3D shape conveyed through texture or shading cues.

MAVERICK Project-Team

3. Research Program

3.1. Introduction

The Maverick project-team aims at producing representations and algorithms for efficient, high-quality computer generation of pictures and animations through the study of four **research problems**:

- *Computer Visualization* where we take as input a large localized dataset and represent it in a way that will let an observer understand its key properties. Visualization can be used for data analysis, for the results of a simulation, for medical imaging data...
- *Expressive Rendering*, where we create an artistic representation of a virtual world. Expressive rendering corresponds to the generation of drawings or paintings of a virtual scene, but also to some areas of computational photography, where the picture is simplified in specific areas to focus the attention.
- *Illumination Simulation*, where we model the interaction of light with the objects in the scene, resulting in a photorealistic picture of the scene. Research include improving the quality and photorealism of pictures, including more complex effects such as depth-of-field or motion-blur. We are also working on accelerating the computations, both for real-time photorealistic rendering and offline, high-quality rendering.
- *Complex Scenes*, where we generate, manage, animate and render highly complex scenes, such as natural scenes with forests, rivers and oceans, but also large datasets for visualization. We are especially interested in interactive visualization of complex scenes, with all the associated challenges in terms of processing and memory bandwidth.

The fundamental research interest of Maverick is first, *understanding* what makes a picture useful, powerful and interesting for the user, and second *designing* algorithms to create and improve these pictures.

3.2. Research approaches

We will address these research problems through three interconnected research approaches:

3.2.1. Picture Impact

Our first research axis deals with the *impact* pictures have on the viewer, and how we can improve this impact. Our research here will target:

- *evaluating user response:* we need to evaluate how the viewers respond to the pictures and animations generated by our algorithms, through user studies, either asking the viewer about what he perceives in a picture or measuring how his body reacts (eye tracking, position tracking).
- *removing artefacts and discontinuities:* temporal and spatial discontinuities perturb viewer attention, distracting the viewer from the main message. These discontinuities occur during the picture creation process; finding and removing them is a difficult process.

3.2.2. Data Representation

The data we receive as input for picture generation is often unsuitable for interactive high-quality rendering: too many details, no spatial organisation... Similarly the pictures we produce or get as input for other algorithms can contain superfluous details.

One of our goals is to develop new data representations, adapted to our requirements for rendering. This includes fast access to the relevant information, but also access to the specific hierarchical level of information needed: we want to organize the data in hierarchical levels, pre-filter it so that sampling at a given level also gives information about the underlying levels. Our research for this axis include filtering, data abstraction, simplification and stylization.

The input data can be of any kind: geometric data, such as the model of an object, scientific data before visualization, pictures and photographs. It can be time-dependent or not; time-dependent data bring an additional level of challenge on the algorithm for fast updates.

3.2.3. Prediction and simulation

Our algorithms for generating pictures require computations: sampling, integration, simulation... These computations can be optimized if we already know the characteristics of the final picture. Our recent research has shown that it is possible to predict the local characteristics of a picture by studying the phenomena involved: the local complexity, the spatial variations, their direction...

Our goal is to develop new techniques for predicting the properties of a picture, and to adapt our image-generation algorithms to these properties, for example by sampling less in areas of low variation.

Our research problems and approaches are all cross-connected. Research on the *impact* of pictures is of interest in three different research problems: *Computer Visualization*, *Expressive rendering* and *Illumination Simulation*. Similarly, our research on *Illumination simulation* will use all three research approaches: impact, representations and prediction.

3.3. Cross-cutting research issues

Beyond the connections between our problems and research approaches, we are interested in several issues, which are present throughout all our research:

sampling is an ubiquitous process occurring in all our application domains, whether photorealistic rendering (*e.g.* photon mapping), expressive rendering (*e.g.* brush strokes), texturing, fluid simulation (Lagrangian methods), etc. When sampling and reconstructing a signal for picture generation, we have to ensure both coherence and homogeneity. By *coherence*, we mean not introducing spatial or temporal discontinuities in the reconstructed signal. By *homogeneity*, we mean that samples should be placed regularly in space and time. For a time-dependent signal, these requirements are conflicting with each other, opening new areas of research.

filtering is another ubiquitous process, occuring in all our application domains, whether in realistic rendering (*e.g.* for integrating height fields, normals, material properties), expressive rendering (*e.g.* for simplifying strokes), textures (through non-linearity and discontinuities). It is especially relevant when we are replacing a signal or data with a lower resolution (for hierarchical representation); this involves filtering the data with a reconstruction kernel, representing the transition between levels.

performance and scalability are also a common requirement for all our applications. We want our algorithms to be usable, which implies that they can be used on large and complex scenes, placing a great importance on scalability. For some applications, we target interactive and real-time applications, with an update frequency between 10 Hz and 120 Hz.

coherence and continuity in space and time is also a common requirement of realistic as well as expressive models which must be ensured despite contradictory requirements. We want to avoid flickering and aliasing.

animation: our input data is likely to be time-varying (*e.g.* animated geometry, physical simulation, time-dependent dataset). A common requirement for all our algorithms and data representation is that they must be compatible with animated data (fast updates for data structures, low latency algorithms...).

3.4. Methodology

Our research is guided by several methodological principles:

Experimentation: to find solutions and phenomenological models, we use experimentation, performing statistical measurements of how a system behaves. We then extract a model from the experimental data.

Validation: for each algorithm we develop, we look for experimental validation: measuring the behavior of the algorithm, how it scales, how it improves over the state-of-the-art... We also compare our algorithms to the exact solution. Validation is harder for some of our research domains, but it remains a key principle for us.

Reducing the complexity of the problem: the equations describing certain behaviors in image synthesis can have a large degree of complexity, precluding computations, especially in real time. This is true for physical simulation of fluids, tree growth, illumination simulation... We are looking for *emerging phenomena* and *phenomenological models* to describe them (see framed box "Emerging phenomena"). Using these, we simplify the theoretical models in a controlled way, to improve user interaction and accelerate the computations.

Transferring ideas from other domains: Computer Graphics is, by nature, at the interface of many research domains: physics for the behavior of light, applied mathematics for numerical simulation, biology, algorithmics... We import tools from all these domains, and keep looking for new tools and ideas.

Develop new fondamental tools: In situations where specific tools are required for a problem, we will proceed from a theoretical framework to develop them. These tools may in return have applications in other domains, and we are ready to disseminate them.

Collaborate with industrial partners: we have a long experience of collaboration with industrial partners. These collaborations bring us new problems to solve, with short-term or medium-term transfer opportunities. When we cooperate with these partners, we have to find *what they need*, which can be very different from *what they want*, their expressed need.

MFX Project-Team

3. Research Program

3.1. Research Program

We focus on the computational aspects of shape modeling and processing for digital fabrication: dealing with shape complexity, revisiting design and customization of existing parts in view of the novel possibilities afforded by AM, and providing a stronger integration between modeling and the capabilities of the target processes.
We tackle on the following challenges:

- develop **novel shape synthesis and shape completion algorithms** that can help users model shapes with features in the scale of microns to meters, while following functional, structural, geometric and fabrication requirements;
- propose methodologies to help *expert* designers **describe shapes** and designs that can later be **customized and adapted** to different use cases;
- develop novel algorithms to **adapt and prepare complex designs** for fabrication in a given technology, including the possibility to modify aspects of the design while preserving its functionality;
- develop novel techniques to **unlock the full potential of fabrication processes**, improving their versatility in terms of feasible shapes as well as their capabilities in terms of accuracy and quality of deposition;
- develop **novel shape representations, data-structures, visualization and interaction techniques** to support the integration of our approaches into a single, unified software framework that covers the full chain from modeling to printing instructions;
- **integrate novel capabilities** enabled by advances in additive manufacturing processes and materials **in the modeling and processing chains**, in particular regarding the use of functional materials (*e.g.* piezoelectric, conductive, shrinkable).

Our approach is to cast a holistic view on the aforementioned challenges, by considering modeling and fabrication as a single, unified process. Thus, the modeling techniques we seek to develop will take into account the geometric constraints imposed by the manufacturing processes (minimal thickness, overhang angles, trapped material) as well as the desired object functionality (rigidity, porosity). To allow for the modeling of complex shapes, and to adapt the same initial design to different technologies, we propose to develop techniques that can automatically synthesize functional details within parts. At the same time, we will explore ways to increase the versatility of the manufacturing processes, through algorithms that are capable of exploiting additional degrees of freedom (*e.g.*, curved layering [21]), can introduce new capabilities (*e.g.*, material mixing [22]) and improve part accuracy (*e.g.*, adaptive slicing [20]).

Our research program is organized along three main research directions. The first one focuses on the automatic synthesis of shapes with intricate multi-scale geometries, that conform to the constraints of additive manufacturing technologies. The second direction considers geometric and algorithmic techniques for the actual fabrication of the modeled object. We aim to further improve the capabilities of the manufacturing processes with novel deposition strategies. The third direction focuses on computational design algorithms to help model parts with gradient of properties, as well as to help customizing existing designs for their reuse.

These three research directions interact strongly, and cross-pollinate: *e.g.*, novel possibilities in manufacturing unlock novel possibilities in terms of shapes that can be synthesized. Stronger synthesis methods allow for further customization.

MIMETIC Project-Team

3. Research Program

3.1. Biomechanics and Motion Control

Human motion control is a highly complex phenomenon that involves several layered systems, as shown in Figure 3 . Each layer of this controller is responsible for dealing with perceptual stimuli in order to decide the actions that should be applied to the human body and his environment. Due to the intrinsic complexity of the information (internal representation of the body and mental state, external representation of the environment) used to perform this task, it is almost impossible to model all the possible states of the system. Even for simple problems, there generally exists an infinity of solutions. For example, from the biomechanical point of view, there are much more actuators (i.e. muscles) than degrees of freedom leading to an infinity of muscle activation patterns for a unique joint rotation. From the reactive point of view there exists an infinity of paths to avoid a given obstacle in navigation tasks. At each layer, the key problem is to understand how people select one solution among these infinite state spaces. Several scientific domains have addressed this problem with specific points of view, such as physiology, biomechanics, neurosciences and psychology.

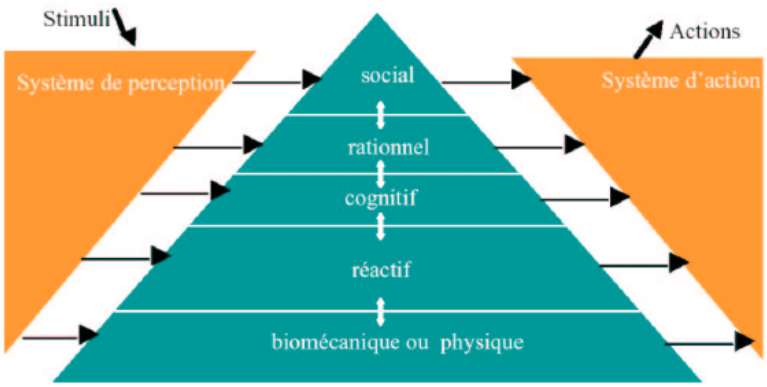


Figure 3. Layers of the motion control natural system in humans.

In biomechanics and physiology, researchers have proposed hypotheses based on accurate joint modeling (to identify the real anatomical rotational axes), energy minimization, force and torques minimization, comfort maximization (i.e. avoiding joint limits), and physiological limitations in muscle force production. All these constraints have been used in optimal controllers to simulate natural motions. The main problem is thus to define how these constraints are composed altogether such as searching the weights used to linearly combine these criteria in order to generate a natural motion. Musculoskeletal models are stereotyped examples for which there exists an infinity of muscle activation patterns, especially when dealing with antagonist muscles. An unresolved problem is to define how to use the above criteria to retrieve the actual activation patterns, while optimization approaches still leads to unrealistic ones. It is still an open problem that will require multidisciplinary skills including computer simulation, constraint solving, biomechanics, optimal control, physiology and neuroscience.

In neuroscience, researchers have proposed other theories, such as coordination patterns between joints driven by simplifications of the variables used to control the motion. The key idea is to assume that instead of controlling all the degrees of freedom, people control higher level variables which correspond to combinations of joint angles. In walking, data reduction techniques such as Principal Component Analysis have shown that lower-limb joint angles are generally projected on a unique plane whose angle in the state space is associated with energy expenditure. Although knowledge exists for specific motions, such as locomotion or grasping, this type of approach is still difficult to generalize. The key problem is that many variables are coupled and it is very difficult to objectively study the behavior of a unique variable in various motor tasks. Computer simulation is a promising method to evaluate such type of assumptions as it enables to accurately control all the variables and to check if it leads to natural movements.

Neuroscience also addresses the problem of coupling perception and action by providing control laws based on visual cues (or any other senses), such as determining how the optical flow is used to control direction in navigation tasks, while dealing with collision avoidance or interception. Coupling of the control variables is enhanced in this case as the state of the body is enriched by the large amount of external information that the subject can use. Virtual environments inhabited with autonomous characters whose behavior is driven by motion control assumptions is a promising approach to solve this problem. For example, an interesting problem in this field is navigation in an environment inhabited with other people. Typically, avoiding static obstacles together with other people displacing into the environment is a combinatory problem that strongly relies on the coupling between perception and action.

One of the main objectives of MimeTIC is to enhance knowledge on human motion control by developing innovative experiments based on computer simulation and immersive environments. To this end, designing experimental protocols is a key point and some of the researchers in MimeTIC have developed this skill in biomechanics and perception-action coupling. Associating these researchers to experts in virtual human simulation, computational geometry and constraints solving enable us to contribute to enhance fundamental knowledge in human motion control.

3.2. Experiments in Virtual Reality

Understanding interactions between humans is challenging because it addresses many complex phenomena including perception, decision-making, cognition and social behaviors. Moreover, all these phenomena are difficult to isolate in real situations, and it is therefore highly complex to understand their individual influence on these human interactions. It is then necessary to find an alternative solution that can standardize the experiments and that allows the modification of only one parameter at a time. Video was first used since the displayed experiment is perfectly repeatable and cut-offs (stop the video at a specific time before its end) allow having temporal information. Nevertheless, the absence of adapted viewpoint and stereoscopic vision does not provide depth information that are very meaningful. Moreover, during video recording session, the real human is acting in front of a camera and not of an opponent. The interaction is then not a real interaction between humans.

Virtual Reality (VR) systems allow full standardization of the experimental situations and the complete control of the virtual environment. It is then possible to modify only one parameter at a time and to observe its influence on the perception of the immersed subject. VR can then be used to understand what information is picked up to make a decision. Moreover, cut-offs can also be used to obtain temporal information about when information is picked up. When the subject can moreover react as in a real situation, his movement (captured in real time) provides information about his reactions to the modified parameter. Not only is the perception studied, but the complete perception-action loop. Perception and action are indeed coupled and influence each other as suggested by Gibson in 1979.

Finally, VR allows the validation of virtual human models. Some models are indeed based on the interaction between the virtual character and the other humans, such as a walking model. In that case, there are two ways to validate it. First, they can be compared to real data (e.g. real trajectories of pedestrians). But such data are not always available and are difficult to get. The alternative solution is then to use VR. The validation of the realism of the model is then done by immersing a real subject in a virtual environment in which a virtual

character is controlled by the model. Its evaluation is then deduced from how the immersed subject reacts when interacting with the model and how realistic he feels the virtual character is.

3.3. Computer animation

Computer animation is the branch of computer science devoted to models for the representation and simulation of the dynamic evolution of virtual environments. A first focus is the animation of virtual characters (behavior and motion). Through a deeper understanding of interactions using VR and through better perceptive, biomechanical and motion control models to simulate the evolution of dynamic systems, the Mimetic team has the ability to build more realistic, efficient and believable animations. Perceptual study also enables us to focus computation time on relevant information (i.e. leading to ensure natural motion from the perceptual points of view) and save time for unperceived details. The underlying challenges are (i) the computational efficiency of the system which needs to run in real-time in many situations, (ii) the capacity of the system to generalise/adapt to new situations for which data was not available or for which models were not defined for, and (iii) the variability of the models, i.e. their ability to handle many body morphologies and generate variations in motions that would be specific to each virtual character.

In many cases, however, these challenges cannot be addressed in isolation. Typically character behaviors also depend on the nature and the topology of the environment they are surrounded by. In essence, a character animation system should also rely on smarter representations of the environments, in order to better perceive the environment, and take contextualised decisions. Hence the animation of virtual characters in our context often requires to be coupled with models to represent the environment, reason, and plan both at a geometric level (can the character reach this location), and at a semantic level (should it use the sidewalk, the stairs, or the road). This represents the second focus. Underlying challenges are the ability to offer a compact, yet precise representation on which efficient path and motion planning can be performed, and on which high-level reasonning can be achieved.

Finally, a third scientific focus tied to the computer animation axis is digital storytelling. Evolved representations of motions and environments enable realistic animations. It is yet equally important to question how these event should be portrayed, when and under which angle. In essence, this means integrating *discourse models* into *story models*, the story representing the sequence of events which occur in a virtual environment, and the discourse representing how this story should be displayed (ie which events to show in which order and with which viewpoint). Underlying challenges are pertained to (i) narrative discourse representations, (ii) projections of the discourse into the geometry, planning camera trajectories and planning cuts between the viewpoints and (iii) means to interactively control the unfolding of the discourse.

By therefore establishing the foundations to build bridges between the high-level narrative structures, the semantic/geometric planning of motions and events, and low-level character animations, the Mimetic team adopts a principled and all-inclusive approach to the animation of virtual characters.

MOEX Project-Team

3. Research Program

3.1. Knowledge representation semantics

We work with semantically defined knowledge representation languages (like description logics, conceptual graphs and object-based languages). Their semantics is usually defined within model theory initially developed for logics.

We consider a language L as a set of syntactically defined expressions (often inductively defined by applying constructors over other expressions). A representation ($o \subseteq L$) is a set of such expressions. It may also be called an ontology. An interpretation function (I) is inductively defined over the structure of the language to a structure called the domain of interpretation (D). This expresses the construction of the "meaning" of an expression in function of its components. A formula is satisfied by an interpretation if it fulfills a condition (in general being interpreted over a particular subset of the domain). A model of a set of expressions is an interpretation satisfying all the expressions. A set of expressions is said consistent if it has at least one model, inconsistent otherwise. An expression (δ) is then a consequence of a set of expressions (o) if it is satisfied by all of their models (noted $o \models \delta$).

The languages dedicated to the semantic web (RDF and OWL) follow that approach. RDF is a knowledge representation language dedicated to the description of resources; OWL is designed for expressing ontologies: it describes concepts and relations that can be used within RDF.

A computer must determine if a particular expression (taken as a query, for instance) is the consequence of a set of axioms (a knowledge base). For that purpose, it uses programs, called provers, that can be based on the processing of a set of inference rules, on the construction of models or on procedural programming. These programs are able to deduce theorems (noted $o \vdash \delta$). They are said to be sound if they only find theorems which are indeed consequences and to be complete if they find all the consequences as theorems.

3.2. Data interlinking with link keys

Vast amounts of RDF data are made available on the web by various institutions providing overlapping information. To be fully exploited, different representations of the same object across various data sets, often using different ontologies, have to be identified. When different vocabularies are used for describing data, it is necessary to identify the concepts they define. This task is called ontology matching and its result is an alignment A, i.e. a set of correspondences $\langle e, r, e' \rangle$ relating entities e and e' of two different ontologies by a particular relation r (which may be equivalence, subsumption, disjointness, etc.) [4].

At the data level, data interlinking is the process of generating links identifying the same resource described in two data sets. Parallel to ontology matching, from two datasets (d and d') it generates a link set, L made of pairs of resource identifier.

We have introduced link keys [4], [1] which extend database keys in a way which is more adapted to RDF and deals with two data sets instead of a single relation. An example of a link key expression is:

$$\{\langle \mathsf{auteur}, \mathsf{creator} \rangle\}\{\langle \mathsf{titre}, \mathsf{title} \rangle\}\ \mathit{linkkey}\ \langle \mathsf{Livre}, \mathsf{Book} \rangle$$

stating that whenever an instance of the class Livre has the same values for the property auteur as an instance of class Book has for the property creator and they share at least one value for their property titre and title, then they denote the same entity. More precisely, a link key is a structure $\langle K^{eq}, K^{in}, C \rangle$ such that:

- K^{eq} and K^{in} are sets of pairs of property expressions;
- C is a pair of class expressions (or a correspondence).

Such a link key holds if and only if for any pair of resources belonging to the classes in correspondence such that the values of their property in K^{eq} are pairwise equal and the values of those in K^{in} pairwise intersect, the resources are the same. Link keys can then be used for finding equal individuals across two data sets and generating the corresponding owl:sameAs links. Link keys take into account the non functionality of RDF data and have to deal with non literal values. In particular, they may use arbitrary properties and class expressions. This renders their discovery and use difficult.

3.3. Experimental cultural knowledge evolution

Cultural evolution considers how culture spreads and evolves with human societies [21]. It applies an an idealised version of the theory of evolution to culture. In computer science, cultural evolution experiments are performed through multi-agent simulation: a society of agents adapts its culture through a precisely defined protocol [16]: agents perform repeatedly and randomly a specific task, called game, and their evolution is monitored. This aims at discovering experimentally the states that agents reach and the properties of these states.

Experimental cultural evolution has been successfully and convincingly applied to the evolution of natural languages [12], [23]. Agents play *language games* and adjust their vocabulary and grammar as soon as they are not able to communicate properly, i.e. they misuse a term or they do not behave in the expected way. It showed its capacity to model various such games in a systematic framework and to provide convincing explanations of linguistic phenomena. Such experiments have shown how agents can agree on a colour coding system or a grammatical case system.

Work has recently been developed for evolving alignments between ontologies. It can be used to repair alignments better than blind logical repair [19], to create alignments based on entity descriptions [13], to learn alignments from dialogues framed in interaction protocols [14], [18], or to correct alignments until no error remains [17][3] and to start with no alignment [2]. Each study provides new insights and opens perspectives.

We adapt this experimental strategy to knowledge representation [3]. Agents use their, shared or private, knowledge to play games and, in case of failure, they use adaptation operators to modify this knowledge. We monitor the evolution of agent knowledge with respect to their ability to perform the game (success rate) and with respect to the properties satisfied by the resulting knowledge itself. Such properties may, for instance, be:

- Agents converge to a common knowledge representation (a convergence property).
- Agents converge towards different but compatible (logically consistent) knowledge (a logical epistemic property), or towards closer knowledge (a metric epistemic property).
- That under the threat of a changing environment, agents that have operators that preserve diverse knowledge recover faster from the changes than those that have operators that converge towards a single representation (a differential property under environment change).

Our goal is to determine which operators are suitable for achieving desired properties in the context of a particular game.

MORPHEO Project-Team

3. Research Program

3.1. Shape and Appearance Modeling

Standard acquisition platforms, including commercial solutions proposed by companies such as Microsoft, 3dMD or 4DViews, now give access to precise 3D models with geometry, e.g. meshes, and appearance information, e.g. textures. Still, state-of-the-art solutions are limited in many respects: They generally consider limited contexts and close setups with typically at most a few meter side lengths. As a result, many dynamic scenes, even a body running sequence, are still challenging situations; They also seldom exploit time redundancy; Additionally, data driven strategies are yet to be fully investigated in the field. The MORPHEO team builds on the Kinovis platform for data acquisition and has addressed these issues with, in particular, contributions on time integration, in order to increase the resolution for both shapes and appearances, on representations, as well as on exploiting recent machine learning tools when modeling dynamic scenes. Our originality lies, for a large part, in the larger scale of the dynamic scenes we consider as well as in the time super resolution strategy we investigate. Another particularity of our research is a strong experimental foundation with the multiple camera Kinovis platforms.

3.2. Dynamic Shape Vision

Dynamic Shape Vision refers to research themes that consider the motion of dynamic shapes, with e.g. shapes in different poses, or the deformation between different shapes, with e.g. different human bodies. This includes for instance shape tracking, shape registration, all these themes being covered by MORPHEO. While progress has been made over the last decade in this domain, challenges remain, in particular due to the required essential task of shape correspondence that is still difficult to perform robustly. Strategies in this domain can be roughly classified into two categories: (i) data driven approaches that learn shape spaces and estimate shapes and their variations through space parameterizations; (ii) model based approaches that use more or less constrained prior models on shape evolutions, e.g. locally rigid structures, to recover correspondences. The MORPHEO team is substantially involved in the second category that leaves more flexibility for shapes that can be modeled, an important feature with the Kinovis platform. The team is anyway also considering the first category with faces and body under clothes modeling, classes of shapes that are more likely to evolve in spaces with reasonable dimensions. The originality of MORPHEO in this axis is to go beyond static shape poses and to consider also the dynamics of shape over several frames when modeling moving shapes, this in particular with shape tracking, animation and, more recently, face registration.

3.3. Inside Shape Vision

Another research axis is concerned with the ability to perceive inside moving shapes. This is a more recent research theme in the MORPHEO team that has gained importance. It was originally the research associated to the Kinovis platform installed in the Grenoble Hospitals. This platform is equipped with two X-ray cameras and ten color cameras, enabling therefore simultaneous vision of inside and outside shapes. We believe this opens a new domain of investigation at the interface between computer vision and medical imaging. Interesting issues in this domain include the links between the outside surface of a shape and its inner parts, especially with the human body. These links are likely to help understanding and modeling human motions. Until now, numerous dynamic shape models, especially in the computer graphic domain, consist of a surface, typically a mesh, bound to a skeletal structure that is never observed in practice but that help anyway parameterizing human motion. Learning more accurate relationships using observations can therefore significantly impact the domain.

3.4. Shape Animation

3D animation is a crucial part of digital media production with numerous applications, in particular in the game and motion picture industry. Recent evolutions in computer animation consider real videos for both the creation and the animation of characters. The advantage of this strategy is twofold: it reduces the creation cost and increases realism by considering only real data. Furthermore, it allows to create new motions, for real characters, by recombining recorded elementary movements. In addition to enable new media contents to be produced, it also allows to automatically extend moving shape datasets with fully controllable new motions. This ability appears to be of great importance with the recent advent of deep learning techniques and the associated need for large learning datasets. In this research direction, we investigate how to create new dynamic scenes using recorded events.

MULTISPEECH Project-Team

3. Research Program

3.1. Beyond black-box supervised learning

This research axis focuses on fundamental, domain-agnostic challenges relating to deep learning, such as the integration of domain knowledge, data efficiency, or privacy preservation. The results of this axis naturally apply in the domains studied in the two other research axes.

3.1.1. Integrating domain knowledge

State-of-the-art methods in speech and audio are based on neural networks trained for the targeted task. This paradigm faces major limitations: lack of interpretability and of guarantees, large data requirements, and inability to generalize to unseen classes or tasks. We intend to research **deep generative models** as a way to learn task-agnostic probabilistic models of audio signals and design inference methods to combine and reuse them for a variety of tasks. We will pursue our investigation of hybrid methods that combine the representational power of deep learning with **statistical signal processing** expertise by leveraging recent optimization techniques for non-convex, non-linear inverse problems. We will also explore the integration of deep learning and **symbolic reasoning** to increase the generalization ability of deep models and to empower researchers/engineers to improve them.

3.1.2. Learning from little/no labeled data

While fully labeled data are costly, unlabeled data are cheap but provide intrinsically less information. **Weakly supervised learning** based on not-so-expensive incomplete and/or noisy labels is a promising middle ground. This entails modeling label noise and leveraging it for unbiased training. Models may depend on the labeler, the spoken context (voice command), or the temporal structure (ambient sound analysis). We will also keep studying **transfer learning** to adapt an expressive (audiovisual) speech synthesizer trained on a given speaker to another speaker for which only neutral voice data has been collected.

3.1.3. Preserving privacy

Some voice technology companies process users' voices in the cloud and store them for training purposes, which raises privacy concerns. We aim to **hide speaker identity** and (some) speaker states and traits from the speech signal, and evaluate the resulting automatic speech/speaker recognition accuracy and subjective quality/intelligibility/identifiability, possibly after removing private words from the training data. We will also explore **semi-decentralized learning** methods for model personalization, and seek to obtain statistical guarantees.

3.2. Speech production and perception

This research axis covers topics related to the production of speech through articulatory modeling and multimodal expressive speech synthesis, and topics related to the perception of speech through the categorization of sounds and prosody in native and in non-native speech.

3.2.1. Articulatory modeling

Articulatory speech synthesis will rely on further 2D and 3D modeling of the vocal tract as well as of the **dynamics of the vocal tract** from real-time MRI data. The prediction of glottis opening will also be considered so as to produce better quality acoustic events for consonants. The **coarticulation model** developed to handle the animation of the visible articulators will be extended to control the face and the tongue. This will help characterize links between the vocal tract and the face, and illustrate inner mouth articulation to learners. The suspension of articulatory movements in stuttering speech will also be studied.

3.2.2. Multimodal expressive speech

The dynamic realism of the animation of the talking head, which has a direct impact on audiovisual intelligibility, will continue to be our goal. Both the **animation** of the lower part of the face relating to speech and of the upper part relating to the facial expression will be considered, and development will continue towards a multilingual talking head. We will investigate further the modeling of **expressivity** both for audio-only and for audiovisual speech synthesis. We will also evaluate the benefit of the talking head in various use cases, including children with language and learning disabilities or deaf people.

3.2.3. Categorization of sounds and prosody

Reading and speaking are basic skills that need to be mastered. Further analysis of schooling experience will allow a better understanding of reading acquisition, especially for children with some language impairment. With respect to L1/L2 language interference [0] , a special focus will be set on the impact of L2 prosody on segmental realizations. Prosody will also be considered for its implication on the structuration of speech communication, including on discourse particles. Moreover, we will experiment the usage of speech technologies for computer assisted language learning in middle and high schools, and, hopefully, also for helping children learning to read.

3.3. Speech in its environment

The themes covered by this research axis correspond to the acoustic environment analysis, to speech enhancement and noise robustness, and to linguistic and semantic processing.

3.3.1. Acoustic environment analysis

Audio scene analysis is key to characterize the environment in which spoken communication may take place. We will investigate audio event detection methods that exploit both strongly/weakly labeled and unlabeled data, operate in real-world conditions, can discover novel events, and provide a semantic interpretation. We will keep working on source localization in the presence of nearby acoustic reflectors. We will also pursue our effort at the interface of **room acoustics** to blindly estimate room properties and develop acoustics-aware signal processing methods. Beyond spoken communication, this has many applications to surveillance, robot audition, building acoustics, and augmented reality.

3.3.2. Speech enhancement and noise robustness

We will pursue **speech enhancement** methods targeting several distortions (echo, reverberation, noise, overlapping speech) for both speech and speaker recognition applications, and extend them to ad-hoc arrays made of the microphones available in our daily life using multi-view learning. We will also continue to explore statistical signal models **beyond the usual zero-mean complex Gaussian model** in the time-frequency domain, e.g., deep generative models of the signal phase. **Robust acoustic modeling** will be achieved by learning domain-invariant representations or performing unsupervised domain adaptation on the one hand, and by extending our uncertainty-aware approach to more advanced (e.g., nongaussian) uncertainty models and accounting for the additional uncertainty due to short utterances on the other hand, with application to speaker and language recognition “in the wild”.

3.3.3. Linguistic and semantic processing

We will seek to address robust speech recognition by exploiting word/sentence embeddings carrying **semantic information** and combining them with acoustical uncertainty to rescore the recognizer outputs. We will also combine semantic content analysis with text obfuscation models (similar to the label noise models to be investigated for weakly supervised training of speech recognition) for the task of detecting and classifying (hateful, aggressive, insulting, ironic, neutral, etc.) **hate speech** in social media.

[0] L1 refers to the speaker’s native language, and L2 to a speaker’s second language, usually learned later as a foreign language

ORPAILLEUR Project-Team

3. Research Program

3.1. Hybrid and Exploratory Knowledge Discovery

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining, data exploration, formal concept analysis, classification, pattern mining, numerical methods in data mining.

Knowledge discovery in databases (KDD) aims at discovering intelligible and reusable patterns in possibly large databases. These patterns can then be interpreted as knowledge units to be reused in knowledge-based systems. From an operational point of view, the KDD process is based on three main steps: (i) selection and preparation of the data, (ii) data mining, (iii) interpretation of the discovered patterns. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. In this view, KDD is an exploratory process similar to "exploratory data analysis".

The KDD process –as implemented in the Orpailleur team– is based on data mining methods which are either symbolic or numerical. Symbolic methods are based on pattern mining (e.g. mining frequent itemsets, association rules, sequences...), Formal Concept Analysis (FCA) and extensions such as Pattern Structures and Relational Concept Analysis (RCA), and redescription mining. Numerical methods are based on Random Forests, Support Vector Machines (SVM), Neural Networks, and probabilistic approaches such as second-order Hidden Markov Models (HMM). Moreover, for being able to deal with complex data, numerical data mining methods can be associated with symbolic methods, for improving applicability and efficiency of knowledge discovery. This is particularly true in classification, where supervised and unsupervised approaches may be combined with benefits.

A main operation in the research work of Orpailleur is "classification", which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. In this way, domain knowledge, when available, can improve and guide the KDD process, materializing the idea of *Knowledge Discovery guided by Domain Knowledge* or KDDK. In KDDK, domain knowledge plays a role at each step of KDD: the discovered patterns can be interpreted as knowledge units and reused for problem-solving activities in knowledge systems, implementing the exploratory process "mining, interpreting, modeling, representing, and reasoning". Then knowledge discovery can be considered as a key task in knowledge engineering (KE), having an impact in various semantic activities, e.g. information retrieval, recommendation, and ontology engineering. In addition, if knowledge discovery can feed knowledge-based systems, in turn, domain knowledge can be used to support the knowledge discovery process.

Finally, life sciences, i.e. agronomy, biology, chemistry, and medicine, are application domains where the Orpailleur team has a very rich experience. The team intends to keep and to extend this experience, paying also more attention to the impact of knowledge discovery in the real world. This should lead to the design of green (sustainable), explainable, and fair data mining systems.

3.2. Text Mining

Keywords: text mining, knowledge discovery from texts, text classification, annotation, ontology engineering from texts.

The objective of a text mining process is to extract useful knowledge units from large collections of texts [71]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a difficult task. A text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a knowledge discovery perspective, text mining aims at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within a knowledge base. The process is roughly similar for text annotation. Text mining is especially useful in the context of semantic web for ontology engineering. In the Orpailleur team, we work on the mining of real-world texts in application domains such as biology and medicine, using numerical and symbolic data mining methods. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.3. Knowledge Systems and Web of Data

Keywords: knowledge engineering, web of data, semantic web, ontology, description logics, classification-based reasoning, case-based reasoning, information retrieval, recommendation.

The web of data constitutes a good platform for experimenting ideas on knowledge engineering (KE) and knowledge discovery. A software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available. This is why domain knowledge and ontologies are of main importance. OWL ("Web Ontology Language" https://www.w3.org/OWL/) is based on description logics (DLs [72]) and is the representation language commonly used for designing ontologies. In OWL, knowledge units are represented by classes having properties and instances. Concepts are organized within a partially ordered set based on a subsumption relation, and the inference services are based on subsumption and classification.

Actually, there are many interconnections between concept lattices in FCA and ontologies, e.g. the partial order underlying an ontology can be supported by a concept lattice. Moreover, a pair of implications within a concept lattice can provide a possible materialization of a concept definition in an ontology. In this way, we study how the web of data, considered as a set of knowledge sources, e.g. DBpedia, Wikipedia, Yago, Freebase, can be mined for guiding the design of a knowledge base, and further, how knowledge discovery techniques can be applied for allowing a better usage of the web of data, e.g. Linked Open Data (LOD) classification and completion.

Then, a part of the research work in Knowledge Engineering is oriented towards knowledge discovery in the web of data, as, with the increased interest in machine processable data, more and more data is now published in RDF (Resource Description Framework) format. Particularly, we are interested in the completeness of the data and their potential to provide concept definitions in terms of necessary and sufficient conditions. We have proposed algorithms based on FCA and Redescription Mining which allow data exploration as well as the discovery of definition (bidirectional implication rules).

PANAMA Project-Team

3. Research Program

3.1. Axis 1: Sparse Models and Representations

3.1.1. Efficient Sparse Models and Dictionary Design for Large-scale Data

Sparse models are at the core of many research domains where the large amount and high-dimensionality of digital data requires concise data descriptions for efficient information processing. Recent breakthroughs have demonstrated the ability of these models to provide concise descriptions of complex data collections, together with algorithms of provable performance and bounded complexity.

A crucial prerequisite for the success of today's methods is the knowledge of a "dictionary" characterizing how to concisely describe the data of interest. Choosing a dictionary is currently something of an "art", relying on expert knowledge and heuristics.

Pre-chosen dictionaries such as wavelets, curvelets or Gabor dictionaries, are based upon stylized signal models and benefit from fast transform algorithms, but they fail to fully describe the content of natural signals and their variability. They do not address the huge diversity underlying modern data much beyond time series and images: data defined on graphs (social networks, internet routing, brain connectivity), vector valued data (diffusion tensor imaging of the brain), multichannel or multi-stream data (audiovisual streams, surveillance networks, multimodal biomedical monitoring).

The alternative to a pre-chosen dictionary is a trained dictionary learned from signal instances. While such representations exhibit good performance on small-scale problems, they are currently limited to low-dimensional signal processing due to the necessary training data, memory requirements and computational complexity. Whether designed or learned from a training corpus, dictionary-based sparse models and the associated methodology fail to scale up to the volume and resolution of modern digital data, for they intrinsically involve difficult linear inverse problems. To overcome this bottleneck, a new generation of efficient sparse models is needed, beyond dictionaries, encompassing the ability to provide sparse and structured data representations as well as computational efficiency. For example, while dictionaries describe low-dimensional signal models in terms of their "synthesis" using few elementary building blocks called atoms, in "analysis" alternatives the low-dimensional structure of the signal is rather "carved out" by a set of equations satisfied by the signal. Linear as well as nonlinear models can be envisioned.

3.1.2. Compressive Learning

A flagship emerging application of sparsity is the paradigm of compressive sensing, which exploits sparse models at the analog and digital levels for the acquisition, compression and transmission of data using limited resources (fewer/less expensive sensors, limited energy consumption and transmission bandwidth, etc.). Besides sparsity, a key pillar of compressive sensing is the use of random low-dimensional projections. Through compressive sensing, random projections have shown their potential to allow drastic dimension reduction with controlled information loss, provided that the projected signal vector admits a sparse representation in some transformed domain. A related scientific domain, where sparsity has been recognized as a key enabling factor, is Machine Learning, where the overall goal is to design statistically founded principles and efficient algorithms in order to infer general properties of large data collections through the observation of a limited number of representative examples. Marrying sparsity and random low-dimensional projections with machine learning shall allow the development of techniques able to efficiently capture and process the information content of large data collections. The expected outcome is a dramatic increase of the impact of sparse models in machine learning, as well as an integrated framework from the signal level (signals and their acquisition) to the semantic level (information and its manipulation), and applications to data sizes and volumes of collections that cannot be handled by current technologies.

3.2. Axis 2: Robust Acoustic Scene Analysis

3.2.1. Compressive Acquisition and Processing of Acoustic Scenes

Acoustic imaging and scene analysis involve acquiring the information content from acoustic fields with a limited number of acoustic sensors. A full 3D+t field at CD quality and Nyquist spatial sampling represents roughly 10^6 microphones/m^3. Dealing with such high-dimensional data requires to drastically reduce the data flow by positioning appropriate sensors, and selecting from all spatial locations the few spots where acoustic sources are active. The main goal is to develop a theoretical and practical understanding of the conditions under which compressive acoustic sensing is both feasible and robust to inaccurate modeling, noisy measures, and partially failing or uncalibrated sensing devices, in various acoustic sensing scenarii. This requires the development of adequate algorithmic tools, numerical simulations, and experimental data in simple settings where hardware prototypes can be implemented.

3.2.2. Robust Audio Source Separation

Audio signal separation consists in extracting the individual sound of different instruments or speakers that were mixed on a recording. It is now successfully addressed in the academic setting of linear instantaneous mixtures. Yet, real-life recordings, generally associated to reverberant environments, remain an unsolved difficult challenge, especially with many sources and few audio channels. Much of the difficulty comes from the combination of (i) complex source characteristics, (ii) sophisticated underlying mixing model and (iii) adverse recording environments. Moreover, as opposed to the "academic" blind source separation task, most applicative contexts and new interaction paradigms offer a variety of situations in which prior knowledge and adequate interfaces enable the design and the use of informed and/or manually assisted source separation methods.

One of the objectives of PANAMA is to instantiate and validate specific instances of audio source separation approaches and to target them to real-world industrial applications, such as 5.1 movie re-mastering, interactive music soloist control and outdoor speech enhancement. Extensions of the framework are needed to achieve real-time online processing, and advanced constraints or probabilistic priors for the sources at hand need to be designed, while paying attention to computational scalability issues.

In parallel to these efforts, expected progress in sparse modeling for inverse problems shall bring new approaches to source separation and modeling, as well as to source localization, which is often an important first step in a source separation workflow.

3.2.3. Robust Audio Source Localization

Audio source localization consists in estimating the position of one or several sound sources given the signals received by a microphone array. Knowing the geometry of an audio scene is often a pre-requisite to perform higher-level tasks such as speaker identification and tracking, speech enhancement and recognition or audio source separation. It can be decomposed into two sub-tasks : (i) compute spatial auditory features from raw audio input and (ii) map these features to the desired spatial information. Robustly addressing both these aspects with a limited number of microphones, in the presence of noise, reverberation, multiple and possibly moving sources remains a key challenge in audio signal processing. The first aspect will be tackled by both advanced statistical and acoustical modeling of spatial auditory features. The second one will be addressed by two complementary approaches. *Physics-driven* approaches cast sound source localization as an inverse problem given the known physics of sound propagation within the considered system. *Data-driven* approaches aim at learning the desired feature-to-source-position mapping using real-world or synthetic training datasets adapted to the problem at hand. Combining these approaches should allow a widening of the notion of source localization, considering problems such as the identification of the directivity or diffuseness of the source as well as some of the boundary conditions of the room. A general perspective is to investigate the relations between the physical structure of the source and the particular structures that can be discovered or enforced in the representations and models used for characterization, localization and separation.

3.3. Axis 3: Large-scale Audio Content Processing and Self-organization

3.3.1. Motif Discovery in Audio Data

Facing the ever-growing quantity of multimedia content, the topic of motif discovery and mining has become an emerging trend in multimedia data processing with the ultimate goal of developing weakly supervised paradigms for content-based analysis and indexing. In this context, speech, audio and music content, offers a particularly relevant information stream from which meaningful information can be extracted to create some form of "audio icons" (key-sounds, jingles, recurrent locutions, musical choruses, etc ...) without resorting to comprehensive inventories of expected patterns.

This challenge raises several fundamental questions that will be among our core preoccupations over the next few years. The first question is the deployment of motif discovery on a large scale, a task that requires extending audio motif discovery approaches to incorporate efficient time series pattern matching methods (fingerprinting, similarity search indexing algorithms, stochastic modeling, etc.). The second question is that of the use and interpretation of the motifs discovered. Linking motif discovery and symbolic learning techniques, exploiting motif discovery in machine learning are key research directions to enable the interpretation of recurring motifs.

On the application side, several use cases can be envisioned which will benefit from motif discovery deployed on a large scale. For example, in spoken content, word-like repeating fragments can be used for several spoken document-processing tasks such as language-independent topic segmentation or summarization. Recurring motifs can also be used for audio summarization of audio content. More fundamentally, motif discovery paves the way for a shift from supervised learning approaches for content description to unsupervised paradigms where concepts emerge from the data.

3.3.2. Structure Modeling and Inference in Audio and Musical Contents

Structuring information is a key step for the efficient description and learning of all types of contents, and in particular audio and musical contents. Indeed, structure modeling and inference can be understood as the task of detecting dependencies (and thus establishing relationships) between different fragments, parts or sections of information content.

A stake of structure modeling is to enable more robust descriptions of the properties of the content and better model generalization abilities that can be inferred from a particular content, for instance via cache models, trigger models or more general graphical models designed to render the information gained from structural inference. Moreover, the structure itself can become a robust descriptor of the content, which is likely to be more resistant than surface information to a number of operations such as transmission, transduction, copyright infringement or illegal use.

In this context, information theory concepts need to be investigated to provide criteria and paradigms for detecting and modeling structural properties of audio contents, covering potentially a wide range of application domains in speech content mining, music modeling or audio scene monitoring.

PERCEPTION Project-Team

3. Research Program

3.1. Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [25], [2]. We thoroughly investigated CMM, provided practical resolution algorithms and studied their convergence properties. We developed several methods for sound localization using two or more microphones [1]. The *Gaussian locally-linear model* (GLLiM) is a partially supervised mixture model that allows to map high-dimensional observations (audio, visual, or concatenations of audio-visual vectors) onto low-dimensional manifolds with a partially known structure [9]. This model is particularly well suited for perception because it encodes both observable and unobservable phenomena. A variant of this model, namely *probabilistic piecewise affine mapping* has also been proposed and successfully applied to the problem of sound-source localization and separation [8]. The European projects HUMAVIPS (2010-2013) coordinated by R. Horaud and EARS (2014-2017), applied audio-visual scene analysis to human-robot interaction.

3.2. Stereoscopic Vision

Stereoscopy is one of the most studied topics in biological and computer vision. Nevertheless, classical approaches of addressing this problem fail to integrate eye/camera vergence. From a geometric point of view, the integration of vergence is difficult because one has to re-estimate the epipolar geometry at every new eye/camera rotation. From an algorithmic point of view, it is not clear how to combine depth maps obtained with different eyes/cameras relative orientations. Therefore, we addressed the more general problem of binocular vision that combines the low-level eye/camera geometry, sensor rotations, and practical algorithms based on global optimization [19], [32]. We studied the link between mathematical and computational approaches to stereo (global optimization and Markov random fields) and the brain plausibility of some of these approaches: indeed, we proposed an original mathematical model for the complex cells in visual-cortex areas V1 and V2 that is based on steering Gaussian filters and that admits simple solutions [20]. This addresses the fundamental issue of how local image structure is represented in the brain/computer and how this structure is used for estimating a dense disparity field. Therefore, the main originality of our work is to address both computational and biological issues within a unifying model of binocular vision. Another equally important problem that still remains to be solved is how to integrate binocular depth maps over time. Recently, we have addressed this problem and proposed a semi-global optimization framework that starts with sparse yet reliable matches and proceeds with propagating them over both space and time. The concept of seed-match propagation has then been extended to TOF-stereo fusion [12].

3.3. Audio Signal Processing

Audio-visual fusion algorithms necessitate that the two modalities are represented in the same mathematical space. Binaural audition allows to extract sound-source localization (SSL) information from the acoustic signals recorded with two microphones. We have developed several methods, that perform sound localization in the temporal and the spectral domains. If a direct path is assumed, one can exploit the *time difference of arrival* (TDOA) between two microphones to recover the position of the sound source with respect to the position of the two microphones. The solution is not unique in this case, the sound source lies onto a 2D manifold. However, if one further assumes that the sound source lies in a horizontal plane, it is then possible to extract the azimuth. We used this approach to predict possible sound locations in order to estimate the direction of a speaker [2]. We also developed a geometric formulation and we showed that with four non-coplanar microphones the azimuth and elevation of a single source can be estimated without ambiguity [1]. We also investigated SSL in the spectral domain. This exploits the filtering effects of the head related transfer function (HRTF): there is a different HRTF for the left and right microphones. The interaural spectral features, namely the ILD (interaural level difference) and IPD (interaural phase difference) can be extracted from the short-time Fourier transforms of the two signals. The sound direction is encoded in these interaural features but it is not clear how to make SSL explicit in this case. We proposed a supervised learning formulation that estimates a mapping from interaural spectral features (ILD and IPD) to source directions using two different setups: audio-motor learning [8] and audio-visual learning [10].

3.4. Visual Reconstruction With Multiple Color and Depth Cameras

For the last decade, one of the most active topics in computer vision has been the visual reconstruction of objects, people, and complex scenes using a multiple-camera setup. The PERCEPTION team has pioneered this field and by 2006 several team members published seminal papers in the field. Recent work has concentrated onto the robustness of the 3D reconstructed data using probabilistic outlier rejection techniques combined with algebraic geometry principles and linear algebra solvers [35]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [33]. The originality of this work was to represent photometric information as a scalar function over a discrete Riemannian manifold, thus *generalizing image analysis to mesh and graph analysis*. Manifold equivalents of local-structure detectors and descriptors were developed [34]. The outcome of this pioneering work has been twofold: the formulation of a new research topic now addressed by several teams in the world, and allowed us to start a three year collaboration with Samsung Electronics. We developed the novel concept of *mixed camera systems* combining high-resolution color cameras with low-resolution depth cameras [21], [17],[16]. Together with our start-up company 4D Views Solutions and with Samsung, we developed the first practical depth-color multiple-camera multiple-PC system and the first algorithms to reconstruct high-quality 3D content [12].

3.5. Registration, Tracking and Recognition of People and Actions

The analysis of articulated shapes has challenged standard computer vision algorithms for a long time. There are two difficulties associated with this problem, namely how to represent articulated shapes and how to devise robust registration and tracking methods. We addressed both these difficulties and we proposed a novel kinematic representation that integrates concepts from robotics and from the geometry of vision. In 2008 we proposed a method that parameterizes the occluding contours of a shape with its intrinsic kinematic parameters, such that there is a direct mapping between observed image features and joint parameters [26]. This deterministic model has been motivated by the use of 3D data gathered with multiple cameras. However, this method was not robust to various data flaws and could not achieve state-of-the-art results on standard dataset. Subsequently, we addressed the problem using probabilistic generative models. We formulated the problem of articulated-pose estimation as a maximum-likelihood with missing data and we devised several tractable algorithms [24], [23]. We proposed several expectation-maximization procedures applied to various articulated shapes: human bodies, hands, etc. In parallel, we proposed to segment and register articulated shapes represented with graphs by embedding these graphs using the spectral properties of graph Laplacians [7]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

PERVASIVE Project-Team

3. Research Program

3.1. Modelling Human Awareness and Undestanding

The objectives of this research area are to develop and refine new computational techniques that improve the reliability and performance of situation models, extend the range of possible application domains, and reduce the cost of developing and maintaining situation models. Important research challenges include developing machine-learning techniques to automatically acquire and adapt situation models through interaction, development of techniques to reason and learn about appropriate behaviors, and the development of new algorithms and data structures for representing situation models.

Pervasive has addressed the following research challenges:

Techniques for learning and adapting situation models: Hand crafting of situation models is currently an expensive process requiring extensive trial and error. We will investigate combination of interactive design tools coupled with supervised and semi-supervised learning techniques for constructing initial, simplified prototype situation models in the laboratory. One possible approach is to explore developmental learning to enrich and adapt the range of situations and behaviors through interaction with users.

Reasoning about actions and behaviors: Constructing systems for reasoning about actions and their consequences is an important open challenge. We will explore integration of planning techniques for operationalizing actions sequences within behaviors, and for constructing new action sequences when faced with unexpected difficulties. We will also investigate reasoning techniques within the situation modeling process for anticipating the consequences of actions, events and phenomena.

Algorithms and data structures for situation models: In recent years, we have experimented with an architecture for situated interaction inspired by work in human factors. This model organises perception and interaction as a cyclic process in which directed perception is used to detect and track entities, verify relations between entities, detect trends, anticipate consequences and plan actions. Each phase of this process raises interesting challenges questions algorithms and programming techniques. We will experiment alternative programming techniques representing and reasoning about situation models both in terms of difficulty of specification and development and in terms of efficiency of the resulting implementation. We will also investigate the use of probabilistic graph models as a means to better accommodate uncertain and unreliable information. In particular, we will experiment with using probabilistic predicates for defining situations, and maintaining likelihood scores over multiple situations within a context. Finally, we will investigate the use of simulation as technique for reasoning about consequences of actions and phenomena.

The challenges in this research area have been addressed through three specific research actions covering situation modelling in homes, learning on mobile devices, and reasoning in critical situations.

3.1.1. Learning Routine patterns of activity in the home.

The objective of this research action is to develop a scalable approach to learning routine patterns of activity in a home using situation models. Information about user actions is used to construct situation models in which key elements are semantic representations of time, place, social role and actions. Activities are encoded as sequences of situations. Recurrent activities are detected as sequences of activities that occur at a specific time and place each day. Recurrent activities provide routines what can be used to predict future actions and anticipate needs and services. An early demonstration has been to construct an intelligent assistant that can respond to and filter communications.

This research action is carried out as part of the doctoral research of Julien Cumin in cooperation with researchers at Orange labs, Meylan. Results are to be published at Ubicomp, Ambient intelligence, Intelligent Environments and IEEE Transactions on System Man and Cybernetics. Julien Cumin will complete and defend his doctoral thesis in 2018.

3.1.2. Learning Patterns of Activity with Mobile Devices

The objective of this research action is to develop techniques to observe and learn recurrent patterns of activity using the full suite of sensors available on mobile devices such as tablets and smart phones. Most mobile devices include seven or more sensors organized in 4 groups: Positioning Sensors, Environmental Sensors, Communications Subsystems, and Sensors for Human-Computer Interaction. Taken together, these sensors can provide a very rich source of information about individual activity.

In this area we explore techniques to observe activity with mobiles devices in order to learn daily patterns of activity. We will explore supervised and semi-supervised learning to construct systems to recognize places and relevant activities. Location and place information, semantic time of day, communication activities, interpersonal interactions, and travel activities (walking, driving, riding public transportation, etc.) are recognized as probabilistic predicates and used to construct situation models. Recurrent sequences of situations will be detected and recorded to provide an ability to predict upcoming situations and anticipate needs for information and services.

Our goal is to develop a theory for building context aware services that can be deployed as part of the mobile applications that companies such as SNCF and RATP use to interact with clients. For example, a current project concerns systems that observe daily travel routines for the Paris region RATP metro and SNCF commuter trains. This system learns individual travel routines on the mobile device without the need to divulge information about personal travel to a cloud based system. The resulting service will consult train and metro schedules to assure that planned travel is feasible and to suggest alternatives in the case of travel disruptions. Similar applications are under discussion for the SNCF inter-city travel and Air France for air travel.

This research action is conducted in collaboration with the Inria Startup Situ8ed. The current objective is to deploy and evaluate a first prototype App during 2017. Techniques will be used commercially by Situ8ed for products to be deployed as early as 2019.

3.1.3. Bibliography

[Brdiczka 07] O. Brdiczka, "Learning Situation Models for Context-Aware Services", Doctoral Thesis of the INPG, 25 may 2007.

[Barraquand 12] R. Barraquand, "Design of Sociable Technologies", Doctoral Thesis of the University Grenoble Alps, 2 Feb 2012.

3.2. Perception of People, Activities and Emotions

Machine perception is fundamental for situated behavior. Work in this area concerns construction of perceptual components using computer vision, acoustic perception, accelerometers and other embedded sensors. These include low-cost accelerometers [Bao 04], gyroscopic sensors and magnetometers, vibration sensors, electromagnetic spectrum and signal strength (wifi, bluetooth, GSM), infrared presence detectors, and bolometric imagers, as well as microphones and cameras. With electrical usage monitoring, every power switch can be used as a sensor [Fogarty 06], [Coutaz 16]. We have developed perceptual components for integrated vision systems that combine a low-cost imaging sensors with on-board image processing and wireless communications in a small, low-cost package. Such devices are increasingly available, with the enabling manufacturing technologies driven by the market for integrated imaging sensors on mobile devices. Such technology enables the use of embedded computer vision as a practical sensor for smart objects.

Research challenges addressed in this area include development of practical techniques that can be deployed on smart objects for perception of people and their activities in real world environments, integration and fusion of information from a variety of sensor modalities with different response times and levels of abstraction, and perception of human attention, engagement, and emotion using visual and acoustic sensors.

Work in this research area will focus on three specific Research Actions

3.2.1. Multi-modal perception and modelling of activities

The objective of this research action is to develop techniques for observing and scripting activities for common household tasks such as cooking and cleaning. An important part of this project involves acquiring annotated multi-modal datasets of activity using an extensive suite of visual, acoustic and other sensors. We are interested in real-time on-line techniques that capture and model full body movements, head motion and manipulation actions as 3D articulated motion sequences decorated with semantic labels for individual actions and activities with multiple RGB and RGB-D cameras.

We have explored the integration of 3D articulated models with appearance based recognition approaches and statistical learning for modeling behaviors. Such techniques provide an important enabling technology for context aware services in smart environments [Coutaz 05], [Crowley 15], investigated by Pervasive Interaction team, as well as research on automatic cinematography and film editing investigated by the Imagine team [Gandhi 13] [Gandhi 14] [Ronfard 14] [Galvane 15]. An important challenge is to determine which techniques are most appropriate for detecting, modeling and recognizing a large vocabulary of actions and activities under different observational conditions.

We explored representations of behavior that encodes both temporal-spatial structure and motion at multiple levels of abstraction. We will further propose parameters to encode temporal constraints between actions in the activity classification model using a combination of higher-level action grammars [Pirsiavash 14] and episodic reasoning [Santofimia 14] [Edwards 14].

We have adapted this work to construct narrative descriptions of cooking activities from ego-centric vision, in cooperation with Remi Ronfard of the Imagine Team of Inria.

3.2.2. Perception with low-cost integrated sensors

In this research action, we will continue work on low-cost integrated sensors using visible light, infrared, and acoustic perception. We will continue development of integrated visual sensors that combine micro-cameras and embedded image processing for detecting and recognizing objects in storage areas. We will combine visual and acoustic sensors to monitor activity at work-surfaces. Low cost real-time image analysis procedures will be designed that acquire and process images directly as they are acquired by the sensor.

Bolometric image sensors measure the Far Infrared emissions of surfaces in order to provide an image in which each pixel is an estimate of surface temperature. Within the European MIRTIC project, Grenoble startup, ULIS has created a relatively low-cost Bolometric image sensor (Retina) that provides small images of 80 by 80 pixels taken from the Far-infrared spectrum. Each pixel provides an estimate of surface temperature. Working with Schneider Electric, engineers in the Pervasive Interaction team had developed a small, integrated sensor that combines the MIRTIC Bolometric imager with a microprocessor for on-board image processing. The package has been equipped with a fish-eye lens so that an overhead sensor mounted at a height of 3 meters has a field of view of approximately 5 by 5 meters. Real-time algorithms have been demonstrated for detecting, tracking and counting people, estimating their trajectories and work areas, and estimating posture.

Many of the applications scenarios for Bolometric sensors proposed by Schneider Electric assume a scene model that assigns pixels to surfaces of the floor, walls, windows, desks or other items of furniture. The high cost of providing such models for each installation of the sensor would prohibit most practical applications. We have recently developed a novel automatic calibration algorithm that determines the nature of the surface under each pixel of the sensor.

Work in this area will continue to develop low-cost real time infrared image sensing, as well as explore combinations of far-infrared images with RGB and RGBD images.

3.2.3. Observing and Modelling Competence and Awareness from Eye-gaze and Emotion

Humans display awareness and emotions through a variety of non-verbal channels. It is increasingly possible to record and interpret such information with available technology. Publicly available software can be used to efficiently detect and track face orientation using web cameras. Concentration can be inferred from changes in pupil size [Kahneman 66]. Observation of Facial Action Units [Ekman 71] can be used to detect both sustained

and instantaneous (micro-expressions) displays of valence and excitation. Heart rate can be measured from the Blood Volume Pulse as observed from facial skin color [Poh 11]. Body posture and gesture can be obtained from low-cost RGB sensors with depth information (RGB+D) [Shotton 13] or directly from images using detectors learned using deep learning [Ramakrishna 14]. Awareness and attention can be inferred from eye-gaze (scan path) and fixation using eye-tracking glasses as well as remote eye tracking devices [Holmqvist 11]. Such recordings can be used to reveal awareness of the current situation and to predict ability to respond effectively to opportunities and threats.

This work is supported by the ANR project CEEGE in cooperation with the department of NeuroCognition of Univ. Bielefeld. Work in this area includes the Doctoral research of Thomas Guntz to be defended in 2019.

3.2.4. Bibliography

[Bao 04] L. Bao, and S. S. Intille. "Activity recognition from user-annotated acceleration data.", IEEE Pervasive computing. Springer Berlin Heidelberg, pp. 1-17, 2004.

[Fogarty 06] J. Fogarty, C. Au and S. E. Hudson. "Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition." In Proceedings of the 19th annual ACM symposium on User interface software and technology, UIST 2006, pp. 91-100. ACM, 2006.

[Coutaz 16] J. Coutaz and J.L. Crowley, A First-Person Experience with End-User Development for Smart Homes. IEEE Pervasive Computing, 15(2), pp. 26-39, 2016.

[Coutaz 05] J. Coutaz, J.L. Crowley, S. Dobson, D. Garlan, "Context is key", Communications of the ACM, 48 (3), pp. 49-53, 2005.

[Crowley 15] J. L. Crowley and J. Coutaz, "An Ecological View of Smart Home Technologies", 2015 European Conference on Ambient Intelligence, Athens, Greece, Nov. 2015.

[Gandhi 13] Vineet Gandhi, Remi Ronfard. "Detecting and Naming Actors in Movies using Generative Appearance Models", Computer Vision and Pattern Recognition, 2013.

[Gandhi 14] Vineet Gandhi, Rémi Ronfard, Michael Gleicher. "Multi-Clip Video Editing from a Single Viewpoint", European Conference on Visual Media Production, 2014

[Ronfard 14] R. Ronfard, N. Szilas. "Where story and media meet: computer generation of narrative discourse". Computational Models of Narrative, 2014.

[Galvane 15] Quentin Galvane, Rémi Ronfard, Christophe Lino, Marc Christie. "Continuity Editing for 3D Animation". AAAI Conference on Artificial Intelligence, Jan 2015.

[Pirsiavash 14] Hamed Pirsiavash , Deva Ramanan, "Parsing Videos of Actions with Segmental Grammars", Computer Vision and Pattern Recognition, pp. 612-619, 2014.

[Edwards 14] C. Edwards. 2014, "Decoding the language of human movement". Commun. ACM 57, 12, pp. 12-14, November 2014.

[Kahneman 66] D. Kahneman and J. Beatty, Pupil diameter and load on memory. Science, 154(3756), pp. 1583-1585, 1966.

[Ekman 71] P. Ekman and W.V. Friesen, Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124, 1971.

[Poh 11] M. Z. Poh, D. J. McDuff, and R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng., 58, pp. 7–11, 2011.

[Shotton 13] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, Real-time human pose recognition in parts from single depth images. Commun. ACM, 56, pp. 116–124, 2013.

[Ramakrishna 14] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell and Y. Sheikh, Pose machines: Articulated pose estimation via inference machines. In European Conference on Computer Vision (ECCV 2016), pp. 33-47, Springer, 2014.

[Cao 17] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), IEEE Press, pp. 1302-1310, July, 2017.

[Holmqvist 11] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, Eye Tracking: A Comprehensive Guide to Methods and Measures, OUP Oxford: Oxford, UK, 2011.

3.3. Sociable Interaction with Smart Objects

Reeves and Nass argue that a social interface may be the truly universal interface [Reeves 98]. Current systems lack ability for social interaction because they are unable to perceive and understand humans or to learn from interaction with humans. One of the goals of the research to be performed in Pervasive Interaction is to provide such abilities.

Work in research area RA3 will demonstrate the use of situation models for sociable interaction with smart objects and companion robots. We will explore the use of situation models as a representation for sociable interaction. Our goal in this research is to develop methods to endow an artificial agent with the ability to acquire social common sense using the implicit feedback obtained from interaction with people. We believe that such methods can provide a foundation for socially polite man-machine interaction, and ultimately for other forms of cognitive abilities. We propose to capture social common sense by training the appropriateness of behaviors in social situations. A key challenge is to employ an adequate representation for social situations.

Knowledge for sociable interaction will be encoded as a network of situations that capture both linguistic and non-verbal interaction cues and proper behavioral responses. Stereotypical social interactions will be represented as trajectories through the situation graph. We will explore methods that start from simple stereotypical situation models and extending a situation graph through the addition of new situations and the splitting of existing situations. An important aspect of social common sense is the ability to act appropriately in social situations. We propose to learn the association between behaviors and social situation using reinforcement learning. Situation models will be used as a structure for learning appropriateness of actions and behaviors that may be chosen in each situation, using reinforcement learning to determine a score for appropriateness based on feedback obtained by observing partners during interaction.

Work in this research area will focus on four specific Research Actions

3.3.1. Moving with people

Our objective in this area is to establish the foundations for robot motions that are aware of human social situation that move in a manner that complies with the social context, social expectations, social conventions and cognitive abilities of humans. Appropriate and socially compliant interactions require the ability for real time perception of the identity, social role, actions, activities and intents of humans. Such perception can be used to dynamically model the current situation in order to understand the situation and to compute the appropriate course of action for the robot depending on the task at hand.

To reach this objective, we propose to investigate three interacting research areas:

- Modeling the context and situation of human activities for motion planning
- Planning and acting in a social context.
- Identifying and modeling interaction behaviors.

In particular, we will investigate techniques that allow a tele-presence robot, such as the BEAM system, to autonomously navigate in crowds of people as may be found at the entry to a conference room, or in the hallway of a scientific meeting.

3.3.2. Understanding and communicating intentions from motion

This research area concerns the communication through motion. When two or more people move as a group, their motion is regulated by implicit rules that signal a shared sense of social conventions and social roles. For example, moving towards someone while looking directly at them signals an intention for engagement. In

certain cultures, subtle rules dictate who passes through a door first or last. When humans move in groups, they implicitly communicate intentions with motion. In this research area, we will explore the scientific literature on proxemics and the social sciences on such movements, in order to encode and evaluate techniques for socially appropriate motion by robots.

3.3.3. Socially aware interaction

This research area concerns socially aware man-machine interaction. Appropriate and socially compliant interaction requires the ability for real time perception of the identity, social role, actions, activities and intents of humans. Such perception can be used to dynamically model the current situation in order to understand the context and to compute the appropriate course of action for the task at hand. Performing such interactions in manner that respects and complies with human social norms and conventions requires models for social roles and norms of behavior as well as the ability to adapt to local social conventions and individual user preferences. In this research area, we will complement research area 3.2 with other forms of communication and interaction, including expression with stylistic face expressions rendered on a tablet, facial gestures, body motions and speech synthesis. We will experiment with use of commercially available tool for spoken language interaction in conjunction with expressive gestures.

3.3.4. Stimulating affection and persuasion with affective devices.

This research area concerns technologies that can stimulate affection and engagement, as well as induce changes in behavior. When acting as a coach or cooking advisor, smart objects must be credible and persuasive. One way to achieve this goal is to express affective feedbacks while interacting. This can be done using sound, light and/or complex moves when the system is composed of actuators.

Research in this area will address 3 questions:

1. How do human perceive affective signals expressed by smart objects (including robots)?
2. How does physical embodiment effect perception of affect by humans?
3. What are the most effective models and tools for animation of affective expression?

Both the physical form and the range of motion have important impact on the ability of a system to inspire affection. We will create new models to propose a generic animation model, and explore the effectiveness of different forms of motion in stimulating affect.

3.3.5. Bibliography

[Reeves 98] B. Reeves and C. Nass, The Media Equation: how People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, 1998.

3.4. Interaction with Pervasive Smart Objects and Displays

Currently, the most effective technologies for new media for sensing, perception and experience are provided by virtual and augmented realities [Van Krevelen 2010]. At the same time, the most effective means to augment human cognitive abilities are provided by access to information spaces such as the world-wide-web using graphical user interfaces. A current challenge is to bring these two media together.

Display technologies continue to decrease exponentially, driven largely by investment in consumer electronics as well as the overall decrease in cost of microelectronics. A consequence has been an increasing deployment of digital displays in both public and private spaces. This trend is likely to accelerate, as new technologies and growth in available communications bandwidth enable ubiquitous low-cost access to information and communications.

The arrival of pervasive displays raises a number of interesting challenges for situated multi-modal interaction. For example:

1. Can we use perception to detect user engagement and identify users in public spaces?
2. Can we replace traditional pointing hardware with gaze and gesture based interaction?
3. Can we tailor information and interaction for truly situated interaction, providing the right information at the right time using the right interaction modality?
4. How can we avoid information overload and unnecessary distraction with pervasive displays?

It is increasingly possible to embed sensors and displays in clothing and ordinary devices, leading to new forms of tangible and wearable interaction with information. This raises challenges such as

1. What are the tradeoffs between large-scale environmental displays and wearable displays using technologies such as e-textiles and pico-projector?
2. How can we manage the tradeoffs between implicit and explicit interaction with both tangible and wearable interaction?
3. How can we determine the appropriate modalities for interaction?
4. How can we make users aware of interaction possibilities without creating distraction?

In addition to display and communications, the continued decrease in microelectronics has also driven an exponential decrease in cost of sensors, actuators, and computing resulting in an exponential growth in the number of smart objects in human environments. Current models for systems organization are based on centralized control, in which a controller or local hub, orchestrates smart objects, generally in connection with cloud computing. This model creates problems with privacy and ownership of information. An alternative is to organize local collections of smart objects to provide distributed services without the use of a centralized controller. The science of ecology can provide an architectural model for such organization.

This approach raises a number of interesting research challenges for pervasive interaction:

1. Can we devise distributed models for multi-modal fusion and interaction with information on heterogeneous devices?
2. Can we devise models for distributed interaction that migrates over available devices as the user changes location and task?
3. Can we manage migration of interaction over devices in a manner that provides seamless immersive interaction with information, services and media?
4. Can we provide models of distributed interaction that conserve the interaction context as services migrate?

Research Actions for Interaction with Pervasive Smart Objects for the period 2017 - 2020 include

3.4.1. Wearable and tangible interaction with smart textiles and wearable projectors

Opportunities in this area result from the emergence of new forms of interactive media using smart objects. We will explore the use of smart objects as tangible interfaces that make it possible to experience and interact with information and services by grasping and manipulating objects. We will explore the use of sensors and actuators in clothing and wearable devices such as gloves, hats and wrist bands both as a means of unobtrusively sensing human intentions and emotional states and as a means of stimulating human senses through vibration and sound. We will explore the new forms of interaction and immersion made possible by deploying interactive displays over large areas of an environment.

3.4.2. Pervasive interaction with ecologies of smart objects in the home

In this research area, we will explore and evaluate interaction with ecologies of smart objects in home environments. We will explore development of a range of smart objects that provide information services, such as devices for Episodic Memory for work surfaces and storage areas, devices to provide energy efficient control of environmental conditions, and interactive media that collect and display information. We propose to develop a new class of socially aware managers that coordinate smart objects and manage logistics in functional areas such as the kitchen, living rooms, closets, bedrooms, bathroom or office.

3.4.3. Bibliography

[Van Krevelen 10] D. W. F. Van Krevelen and R. Poelman, A survey of augmented reality technologies, applications and limitations. International Journal of Virtual Reality, 9(2), 1-20, 2010

PETRUS Project-Team

3. Research Program

3.1. Research Program

To tackle the challenge introduced above, we identify three main lines of research:

- (Axis 1) Personal cloud server architectures. Based on the intuition that user control, security and privacy are key properties in the definition of trusted personal cloud solutions, our objective is to propose new architectures (encompassing both software and hardware aspects) for secure personal cloud data management and formally prove important bricks of the architecture. We also focus in this axis on administration models and their enforcement in relation to the architecture of the system, so that the exclusive control of a non expert individual can be ensured.
- (Axis 2) Global query evaluation. The goal of this line of research is to provide capabilities for crossing data belonging to multiple individuals (e.g., performing statistical queries over personal data, computing queries on social graphs or organizing participatory data collection) in a fully decentralized setting while providing strong and personalized privacy guarantees. This means proposing new secure distributed database indexing models and query processing strategies. In addition, we concentrate on locally ensuring to each participant the good behaviour of the processing, such that no collective results can be produced if privacy conditions are not respected by other participants.
- (Axis 3) Economic, legal and societal issues. This research axis is more transverse and entails multidisciplinary research, addressing the links between economic, legal, societal and technological aspects. We will follow here a multi-disciplinary approach based on a 3-step methodology: i) identifying important common issues related to privacy and to the exploitation of personal data; ii) characterizing their dimensions in all relevant disciplines and jointly study their entanglement; iii) validating the proposed analysis, models and trade-offs thanks to in vivo experiments.

These contributions will also rely on tools (algorithms, protocols, proofs, etc.) from other communities, namely security (cryptography, secure multiparty computations, formal methods, differential privacy, etc.) and distributed systems (distributed hash tables, gossip protocols, etc.). Beyond the research actions, we structure our software activity around a single common platform (rather than isolated demonstrators), integrating our main research contributions, called PlugDB. This platform is the cornerstone to help validating our research results through accurate performance measurements on a real platform, a common practice in the DB community, and target the best conferences. It is also a strong vector to federate the team, simplify the bootstrapping of new PhD or master students, conduct multi-disciplinary research and open the way to industrial collaborations and technological transfers.

POTIOC Project-Team

3. Research Program

3.1. Research Program

To achieve our overall objective, we follow two main research axes, plus one transverse axis, as illustrated in Figure 2 .

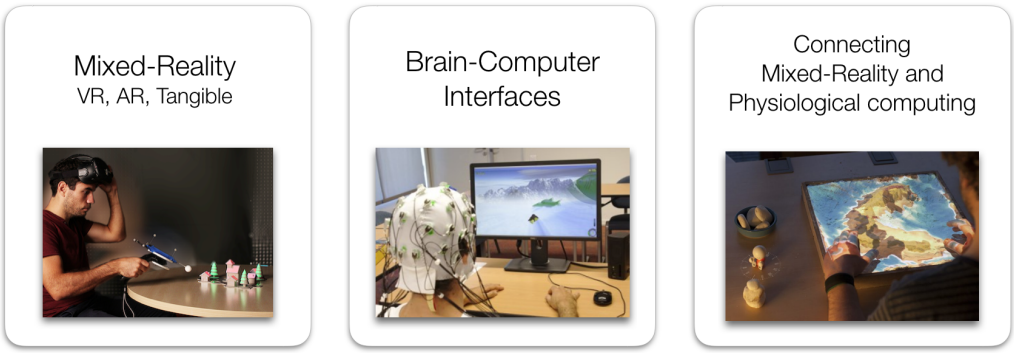


Figure 2. Main research axes of Potioc.

In the first axis dedicated to **Interaction in Mixed-Reality spaces**, we explore interaction paradigms that encompass virtual and/or physical objects. We are notably interested in hybrid environments that co-locate virtual and physical spaces, and we also explore approaches that allow one to move from one space to the other.

The second axis is dedicated to **Brain-Computer Interfaces (BCI)**, i.e., systems enabling user to interact by means of brain activity only. We target BCI systems that are reliable and accessible to a large number of people. To do so, we work on brain signal processing algorithms as well as on understanding and improving the way we train our users to control these BCIs.

Finally, in the **transverse** axis, we explore new approaches that involve both mixed-reality and neuro-physiological signals. In particular, tangible and augmented objects allow us to explore interactive physical visualizations of human inner states. Physiological signals also enable us to better assess user interaction, and consequently, to refine the proposed interaction techniques and metaphors.

From a methodological point of view, for these three axes, we work at three different interconnected levels. The first level is centered on the human sensori-motor and cognitive abilities, as well as user strategies and preferences, for completing interaction tasks. We target, in a fundamental way, a better understanding of humans interacting with interactive systems. The second level is about the creation of interactive systems. This notably includes development of hardware and software components that will allow us to explore new input and output modalities, and to propose adapted interaction techniques. Finally, in a last higher level, we are interested in specific application domains. We want to contribute to the emergence of new applications and usages, with a societal impact.

RAINBOW Project-Team

3. Research Program

3.1. Main Vision

The vision of Rainbow (and foreseen applications) calls for several general scientific challenges: (i) high-level of autonomy for complex robots in complex (unstructured) environments, (ii) forward interfaces for letting an operator giving high-level commands to the robot, (iii) backward interfaces for informing the operator about the robot 'status', (iv) user studies for assessing the best interfacing, which will clearly depend on the particular task/situation. Within Rainbow we plan to tackle these challenges at different levels of depth:

- the **methodological and algorithmic side** of the sought human-robot interaction will be the **main focus** of Rainbow. Here, we will be interest in advancing the state-of-the-art in sensor-based online planning, control and manipulation for mobile/fixed robots. For instance, while classically most control approaches (especially those sensor-based) have been essentially *reactive*, we believe that less myopic strategies based on online/reactive trajectory optimization will be needed for the future Rainbow activities. The core ideas of Model-Predictive Control approaches (also known as Receding Horizon) or, in general, numerical optimal control methods will play a role in the Rainbow activities, for allowing the robots to reason/plan over some future time window and better cope with constraints. We will also consider extending classical sensor-based motion control/manipulation techniques to more realistic scenarios, such as deformable/flexible objects ("**Advanced Sensor-based Control**" axis). Finally, it will also be important to spend research efforts into the field of *Optimal Sensing*, in the sense of generating (again) trajectories that can optimize the state estimation problem in presence of scarce sensory inputs and/or non-negligible measurement and process noises, especially true for the case of mobile robots ("**Optimal and Uncertainty-Aware Sensing**" axis). We also aim at addressing the case of coordination between a single human user and multiple robots where, clearly, as explained the autonomy part plays even a more crucial role (no human can control multiple robots at once, thus a high degree of autonomy will be required by the robot group for executing the human commands);
- the **interfacing side** will also be a focus of the Rainbow activities. As explained above, we will be interested in both the *forward* (human $\rightarrow$ robot) and *backward* (robot $\rightarrow$ human) interfaces. The forward interface will be mainly addressed from the *algorithmic* point of view, i.e., how to map the few degrees of freedom available to a human operator (usually in the order of 3–4) into complex commands for the controlled robot(s). This mapping will typically be mediated by an "AutoPilot" onboard the robot(s) for autonomously assessing if the commands are feasible and, if not, how to least modify them ("**Advanced Sensor-based Control**" axis).
  The backward interface will, instead, mainly consist of a visual/haptic feedback for the operator. Here, we aim at exploiting our expertise in using force cues for informing an operator about the status of the remote robot(s). However, the sole use of classical *grounded* force feedback devices (e.g., the typical force-feedback joysticks) will not be enough due to the different kinds of information that will have to be provided to the operator. In this context, the recent interest in the use of *wearable* haptic interfaces is very interesting and will be investigated in depth (these include, e.g., devices able to provide vibro-tactile information to the fingertips, wrist, or other parts of the body). The main challenges in these activities will be the mechanical conception (and construction) of suitable wearable interfaces for the tasks at hand, and in the generation of force cues for the operator: the force cues will be a (complex) function of the robot state, therefore motivating research in algorithms for mapping the robot state into a few variables (the force cues) ("**Haptics for Robotics Applications**" axis);

- the **evaluation side** that will assess the proposed interfaces with some user studies, or acceptability studies by human subjects. Although this activity **will not** be a main focus of Rainbow (complex user studies are beyond the scope of our core expertise), we will nevertheless devote some efforts into having some reasonable level of user evaluations by applying standard statistical analysis based on psychophysical procedures (e.g., randomized tests and Anova statistical analysis). This will be particularly true for the activities involving the use of smart wheelchairs, which are intended to be used by human users *and* operate inside human crowds. Therefore, we will be interested in gaining some level of understanding of how semi-autonomous robots (a wheelchair in this example) can predict the human intention, and how humans can react to a semi-autonomous mobile robot.

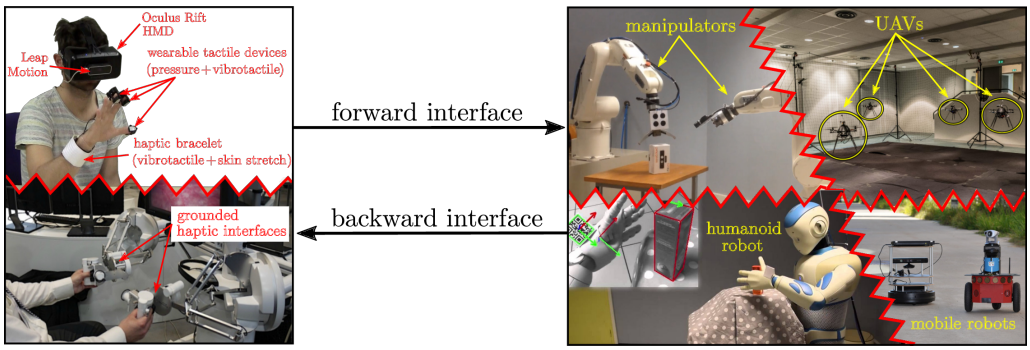


Figure 1. An illustration of the prototypical activities foreseen in Rainbow in which a human operator is in partial (and high-level) control of single/multiple complex robots performing semi-autonomous tasks

Figure 1 depicts in an illustrative way the *prototypical* activities foreseen in Rainbow. On the righthand side, complex robots (dual manipulators, humanoid, single/multiple mobile robots) need to perform some task with high degree of autonomy. On the lefthand side, a human operator gives some high-level commands and receives a visual/haptic feedback aimed at informing her/him at best of the robot status. Again, the main challenges that Rainbow will tackle to address these issues are (in order of relevance): (i) methods and algorithms, mostly based on first-principle modeling and, when possible, on numerical methods for online/reactive trajectory generation, for enabling the robots with high autonomy; (ii) design and implementation of visual/haptic cues for interfacing the human operator with the robots, with a special attention to novel combinations of grounded/ungrounded (wearable) haptic devices; (iii) user and acceptability studies.

3.2. Main Components

Hereafter, a summary description of the four axes of research in Rainbow.

3.2.1. Optimal and Uncertainty-Aware Sensing

Future robots will need to have a large degree of autonomy for, e.g., interpreting the sensory data for accurate estimation of the robot and world state (which can possibly include the human users), and for devising motion plans able to take into account many constraints (actuation, sensor limitations, environment), including also the state estimation accuracy (i.e., how well the robot/environment state can be reconstructed from the sensed data). In this context, we will be particularly interested in (i) devising trajectory optimization strategies able to maximize some norm of the information gain gathered along the trajectory (and with the available sensors). This can be seen as an instance of Active Sensing, with the main focus on *online/reactive* trajectory optimization strategies able to take into account several requirements/constraints (sensing/actuation limitations, noise characteristics). We will also be interested in the coupling between optimal sensing and

concurrent execution of additional tasks (e.g., navigation, manipulation). (ii) Formal methods for guaranteeing the accuracy of localization/state estimation in mobile robotics, mainly exploiting tools from interval analysis. The interest in these methods is their ability to provide possibly conservative but guaranteed accuracy bounds on the best accuracy one can obtain with the given robot/sensor pair, and can thus be used for planning purposes of for system design (choice of the best sensor suite for a given robot/task). (iii) Localization/tracking of objects with poor/unknown or deformable shape, which will be of paramount importance for allowing robots to estimate the state of "complex objects" (e.g., human tissues in medical robotics, elastic materials in manipulation) for controlling its pose/interaction with the objects of interest.

3.2.2. Advanced Sensor-based Control

One of the main competences of the previous Lagadic team has been, generally speaking, the topic of *sensor-based control*, i.e., how to exploit (typically onboard) sensors for controlling the motion of fixed/ground robots. The main emphasis has been in devising ways to directly couple the robot motion with the sensor outputs in order to invert this mapping for driving the robots towards a configuration specified as a desired sensor reading (thus, directly in sensor space). This general idea has been applied to very different contexts: mainly standard vision (from which the Visual Servoing keyword), but also audio, ultrasound imaging, and RGB-D.

Use of sensors for controlling the robot motion will also clearly be a central topic of the Rainbow team too, since the use of (especially onboard) sensing is a main characteristics of any future robotics application (which should typically operate in unstructured environments, and thus mainly rely on its own ability to sense the world). We then naturally aim at making the best out of the previous Lagadic experience in sensor-based control for proposing new advanced ways of exploiting sensed data for, roughly speaking, controlling the motion of a robot. In this respect, we plan to work on the following topics: (i) "direct/dense methods" which try to directly exploit the raw sensory data in computing the control law for positioning/navigation tasks. The advantages of these methods is the need for little data pre-processing which can minimize feature extraction errors and, in general, improve the overall robustness/accuracy (since all the available data is used by the motion controller); (ii) sensor-based interaction with objects of unknown/deformable shapes, for gaining the ability to manipulate, e.g., flexible objects from the acquired sensed data (e.g., controlling online a needle being inserted in a flexible tissue); (iii) sensor-based model predictive control, by developing *online/reactive* trajectory optimization methods able to plan feasible trajectories for robots subjects to sensing/actuation constraints with the possibility of (onboard) sensing for continuously replanning (over some future time horizon) the optimal trajectory. These methods will play an important role when dealing with complex robots affected by complex sensing/actuation constraints, for which pure reactive strategies (as in most of the previous Lagadic works) are not effective. Furthermore, the coupling with the aforementioned optimal sensing will also be considered; (iv) multi-robot decentralised estimation and control, with the aim of devising again sensor-based strategies for groups of multiple robots needing to maintain a formation or perform navigation/manipulation tasks. Here, the challenges come from the need of devising "simple" decentralized and scalable control strategies under the presence of complex sensing constraints (e.g., when using onboard cameras, limited fov, occlusions). Also, the need of locally estimating global quantities (e.g., common frame of reference, global property of the formation such as connectivity or rigidity) will also be a line of active research.

3.2.3. Haptics for Robotics Applications

In the envisaged *shared* cooperation between human users and robots, the typical sensory channel (besides vision) exploited to inform the human users is most often the force/kinesthetic one (in general, the sense of touch and of applied forces to the human hand or limbs). Therefore, a part of our activities will be devoted to study and advance the use of *haptic* cueing algorithms and interfaces for providing a feedback to the users during the execution of some shared task. We will consider: (i) multi-modal haptic cueing for general teleoperation applications, by studying how to convey information through the kinesthetic and cutaneous channels. Indeed, most haptic-enabled applications typically only involve kinesthetic cues, e.g., the forces/torques that can be felt by grasping a force-feedback joystick/device. These cues are very informative about, e.g., preferred/forbidden motion directions, but are also inherently limited in their resolution since the kinesthetic channel can easily become overloaded (when too much information is compressed in a single

cue). In recent years, the arise of novel cutaneous devices able to, e.g., provide vibro-tactile feedback on the fingertips or skin, has proven to be a viable solution to *complement* the classical kinesthetic channel. We will then study how to combine these two sensory modalities for different prototypical application scenarios, e.g., 6-dof teleoperation of manipulator arms, virtual fixtures approaches, and remote manipulation of (possibly deformable) objects; (ii) in the particular context of medical robotics, we plan to address the problem of providing haptic cues for typical medical robotics tasks, such as semi-autonomous needle insertion and robot surgery by exploring the use of kinesthetic feedback for rendering the mechanical properties of the tissues, and vibrotactile feedback for providing with guiding information about pre-planned paths (with the aim of increasing the usability/acceptability of this technology in the medical domain); (iii) finally, in the context of multi-robot control we would like to explore how to use the haptic channel for providing information about the status of *multiple* robots executing a navigation or manipulation task. In this case, the problem is (even more) how to map (or compress) information about many robots into a few haptic cues. We plan to use specialized devices, such as actuated exoskeleton gloves able to provide cues to each fingertip of a human hand, or to resort to "compression" methods inspired by the *hand postural synergies* for providing coordinated cues representative of a few (but complex) motions of the multi-robot group, e.g., coordinated motions (translations/expansions/rotations) or collective grasping/transporting.

3.2.4. Shared Control of Complex Robotics Systems

This final and main research axis will exploit the **methods, algorithms and technologies** developed in the previous axes for realizing applications involving complex semi-autonomous robots operating in complex environments together with human users. The *leitmotiv* is to realize advanced *shared control* paradigms, which essentially aim at blending robot autonomy and user's intervention in an optimal way for exploiting the best of both worlds (robot accuracy/sensing/mobility/strength and human's cognitive capabilities). A common theme will be the issue of where to "draw the line" between robot autonomy and human intervention: obviously, there is no general answer, and any design choice will depend on the particular task at hand and/or on the technological/algorithmic possibilities of the robotic system under consideration.

A *prototypical* envisaged application, exploiting and combining the previous three research axes, is as follows: a complex robot (e.g., a two-arm system, a humanoid robot, a multi-UAV group) needs to operate in an environment exploiting its onboard sensors (in general, vision as the main exteroceptive one) and deal with many constraints (limited actuation, limited sensing, complex kinematics/dynamics, obstacle avoidance, interaction with difficult-to-model entities such as surrounding people, and so on). The robot must then possess a quite large autonomy for interpreting and exploiting the sensed data in order to estimate its own state and the environment one ("**Optimal and Uncertainty-Aware Sensing**" axis), and for planning its motion in order to fulfil the task (e.g., navigation, manipulation) by coping with all the robot/environment constraints. Therefore, advanced control methods able to exploit the sensory data at its most, and able to cope *online* with constraints in an optimal way (by, e.g., continuously replanning and predicting over a future time horizon) will be needed ("**Advanced Sensor-based Control**" axis), with a possible (and interesting) coupling with the sensing part for optimizing, at the same time, the state estimation process. Finally, a human operator will typically be in charge of providing high-level commands (e.g., where to go, what to look at, what to grasp and where) that will then be autonomously executed by the robot, with possible local modifications because of the various (local) constraints. At the same time, the operator will also receive *online* visual-force cues informative of, in general, how well her/his commands are executed and if the robot would prefer or suggest other plans (because of the local constraints that are not of the operator's concern). This information will have to be visually and haptically rendered with an optimal combination of cues that will depend on the particular application ("**Haptics for Robotics Applications**" axis).

RITS Project-Team

3. Research Program

3.1. Vehicle guidance and autonomous navigation

Participants: Mohammad Abualhoul, Pranav Agarwal, Said Alexander Alvarado Marin, Sylia Baraka, Pierre de Beaucorps, Fares Bessam, Pierre Bourre, Raoul de Charette, Carlos Flores, Farouk Ghallabi, Manuel Gonzalez, Maximilian Jaritz, Manohar Kv, Imane Mahtout, Kathia Melbouci, Kaouther Messaoud, Fawzi Nashashibi, Fabio Pizzati, Renaud Poncelet, Danut Ovidiu Pop, Luis Roldao, Anne Verroust-Blondet, Leonardo Ward, Itheri Yahiaoui.

There are three basic ways to improve the safety of road vehicles and these ways are all of interest to the project-team. The first way is to assist the driver by giving him better information and warning. The second way is to take over the control of the vehicle in case of mistakes such as inattention or wrong command. The third way is to completely remove the driver from the control loop.

All three approaches rely on information processing. Only the last two involve the control of the vehicle with actions on the actuators, which are the engine power, the brakes and the steering. The research proposed by the project-team is focused on the following elements:

- perception of the environment,
- planning of the actions,
- real-time control.

3.1.1. Perception of the road environment

Participants: Raoul de Charette, Maximilian Jaritz, Farouk Ghallabi, Manohar Kv, Kaouther Messaoud, Fawzi Nashashibi, Fabio Pizzati, Danut Ovidiu Pop, Luis Roldao, Anne Verroust-Blondet, Itheri Yahiaoui.

Either for driver assistance or for fully automated guided vehicle purposes, the first step of any robotic system is to perceive the environment in order to assess the situation around itself. Proprioceptive sensors (accelerometer, gyrometer,...) provide information about the vehicle by itself such as its velocity or lateral acceleration. On the other hand, exteroceptive sensors, such as video camera, laser or GPS devices, provide information about the environment surrounding the vehicle or its localization. Obviously, fusion of data with various other sensors is also a focus of the research.
The following topics are already validated or under development in our team:

- relative ego-localization with respect to the infrastructure, i.e. lateral positioning on the road can be obtained by mean of vision (lane markings) and the fusion with other devices (e.g. GPS);
- global ego-localization by considering GPS measurement and proprioceptive information, even in case of GPS outage;
- road detection by using lane marking detection and navigable free space;
- detection and localization of the surrounding obstacles (vehicles, pedestrians, animals, objects on roads, etc.) and determination of their behavior can be obtained by the fusion of vision, laser or radar based data processing;
- simultaneous localization and mapping as well as mobile object tracking using laser-based and stereovision-based (SLAMMOT) algorithms.

Scene understanding is a large perception problem. In this research axis we have decided to use only computer vision as cameras have evolved very quickly and can now provide much more precise sensing of the scene, and even depth information. Two types of hardware setups were used, namely: monocular vision or stereo vision to retrieve depth information which allow extracting geometry information.

We have initiated several works:

- estimation of the ego motion using monocular scene flow. Although in the state of the art most of the algorithms use a stereo setup, researches were conducted to estimate the ego-motion using a novel approach with a strong assumption.
- bad weather conditions evaluations. Most often all computer vision algorithms work under a transparent atmosphere assumption which assumption is incorrect in the case of bad weather (rain, snow, hail, fog, etc.). In these situations the light ray are disrupted by the particles in suspension, producing light attenuation, reflection, refraction that alter the image processing.
- deep learning for object recognition. New works are being initiated in our team to develop deep learning recognition in the context of heterogeneous data.
- deep learning for vehicle motion prediction.

3.1.2. Planning and executing vehicle actions

Participants: Pierre de Beaucorps, Carlos Flores, Imane Mahtout, Fawzi Nashashibi, Renaud Poncelet, Anne Verroust-Blondet.

From the understanding of the environment, thanks to augmented perception, we have either to warn the driver to help him in the control of his vehicle, or to take control in case of a driverless vehicle. In simple situations, the planning might also be quite simple, but in the most complex situations we want to explore, the planning must involve complex algorithms dealing with the trajectories of the vehicle and its surroundings (which might involve other vehicles and/or fixed or moving obstacles). In the case of fully automated vehicles, the perception will involve some map building of the environment and obstacles, and the planning will involve partial planning with periodical recomputation to reach the long term goal. In this case, with vehicle to vehicle communications, what we want to explore is the possibility to establish a negotiation protocol in order to coordinate nearby vehicles (what humans usually do by using driving rules, common sense and/or non verbal communication). Until now, we have been focusing on the generation of geometric trajectories as a result of a maneuver selection process using grid-based rating technique or fuzzy technique. For high speed vehicles, Partial Motion Planning techniques we tested, revealed their limitations because of the computational cost. The use of quintic polynomials we designed, allowed us to elaborate trajectories with different dynamics adapted to the driver profile. These trajectories have been implemented and validated in the JointSystem demonstrator of the German Aerospace Center (DLR) used in the European project HAVEit, as well as in RITS's electrical vehicle prototype used in the French project ABV. HAVEit was also the opportunity for RITS to take in charge the implementation of the Co-Pilot system which processes perception data in order to elaborate the high level command for the actuators. These trajectories were also validated on RITS's cybercars. However, for the low speed cybercars that have pre-defined itineraries and basic maneuvers, it was necessary to develop a more adapted planning and control system. Therefore, we have developed a nonlinear adaptive control for automated overtaking maneuver using quadratic polynomials and Lyapunov function candidate and taking into account the vehicles kinematics. For the global mobility systems we are developing, the control of the vehicles includes also advanced platooning, automated parking, automated docking, etc. For each functionality a dedicated control algorithm was designed (see publication of previous years).

3.2. Mobile wireless communications for vehicular networks

Participants: Gérard Le Lann, Mohammad Abualhoul, Fawzi Nashashibi.

Wireless communications are expected to play an essential role in ensuring road safety, road efficiency, and driving comfort o. Road safety applications often require relatively short response time and reliable information exchange between neighboring vehicles and road-side units in any road density condition. Because of the performance of the existing radio communications technology largely degrades with the increase of the traffic density, the challenge of designing wireless communications solution suitable for safety applications is enabling reliable communications in highly dense scenarios.

To investigate this open problem and trade-off situations, RITS has been working on medium access control design for the IEEE 802.11p radio communication and the deployment of supportive solutions such as visible light communications and testing the use-cases for extreme traffic conditions and highly dense scenarios. The works have been carried out considering the vehicle behavior such as autonomous and connected vehicles merging, sharing, and convoy forming as platoon scenarios with considering the hard-safety requirements.

Unlike many of the road safety applications, the applications regarding road efficiency and comfort of road users, often require connectivity to the Internet. Based on our expertise in both Internet-based communications in the mobility context and in ITS, we are investigating the use of IPv6 (Internet Protocol version 6 which is going to replace the current version, IPv4, IoT) for vehicular communications, in a combined architecture supporting both V2V and V2I.

Communication contributions at RITS team have been working on channel modeling for both radio and visible light communications, and design of communications mechanisms, especially for security, service discovery, multicast, and Geo-Cast message delivery, and access point selection.

RITS-team has one of the latest certified standard communication hardware and tools supported by the partnership with the YoGoKo Company. All platforms (connected and autonomous vehicles) are equipped with state-of-art communication units On-Board-Units (OBU), where the Rocquencourt site equipped with two stationary Road-Side-Units (RSU) enabling all kind of tests and projects requirements

Below follows a more detailed description of the related research issues.

3.2.1. Regulation study for interoperability tests for cooperative driving

Participants: Mohammad Abualhoul, Fawzi Nashashibi.

The technological advances of autonomous and connected road vehicles have been shown an accelerating pace in the recent years. On the other hand, the regulations for autonomous, or driverless, road vehicles across Europe still deserve much attention and discussion

Therefore, RITS-Inria team plays a key element in one of the European demonstration-based projects (AUTOC-ITS), which aims to contribute to the regulation study for interoperability in the adoption of autonomous driving in European urban nodes. The regulation study done by RITS team and project partners meant to conduct a deployment of Cooperative Intelligent Transport Systems (C-ITS) in Europe by enhancing interoperability for autonomous vehicles [29]. The project activities and RITS contributions will also boost the role of C-ITS as the primary catalyst for any future implementation of autonomous driving scenarios in Europe. The final demonstration of different European partners will require the implementation and preparations of three pilots sites in three major European cities: Paris, Madrid, and Lisbon. Pilot locations in these major cities are chosen to be located along the European Atlantic Corridor for interoperability evaluation.

RITS-Inria is coordinating the French contribution by evaluating the deployment of C-ITS services in the A13-Paris, which belongs to the French part of the Atlantic Corridor.

Team Core contributions:

- Provide up to date feedback to contribute to the present EU and international regulations on autonomous vehicles.
- Build and evaluate the pilots experimentally by deploying fully autonomous vehicles and a Cooperative Intelligent Transport Systems (C-ITS).
- Define and evaluate a safety autonomous driving services, such as:
  - Roadworks warning.
  - Weather conditions.
  - Other hazardous notifications.
- Define and perform communication interoperability tests between deferent partners for different scenarios, messaging and hardware to ensure the compatibility in using the IEEE 802.11p standard.
- Study the extension of the results on large-scale deployment in other European countries.
- Contribute to the European standards organizations such as C-Roads, C-ITS platforms.

AUTOC-ITS project brings the road authorities from France, Span, and Portugal (DGT, ANSR, SANEF) and C-ITS experts from research institutes and universities (Inria, INDRA, UPM, UC, IPN) to carry out a cooperative work and contributes to the C-ITS Platform by bringing answers to the field of automation driving.

3.2.2. V2X radio communications for road safety applications

Participants: Mohammad Abualhoul, Fawzi Nashashibi.

The development work and generating proper components to facilitate communication requirements and to be deployed in different projects scenarios is one of the main ongoing activities by all RITS team members.

There are continuous activities on both theoretical modeling and experimental evaluation of the radio channel characteristics in vehicular networks, especially the radio quality, channel congestion, load allocations, congestion, and bandwidth availability.

Based on our previous expertise and studies, we develop mechanisms for efficient and reliable V2X communications, access point selection, handover algorithms which are especially dedicated to road safety and autonomous driving applications.

3.2.3. Cyberphysical constructs and mobile communications for fully automated networked vehicles

Participant: Gérard Le Lann.

Intelligent vehicular networks (IVNs) are constituents of ITS. IVNs range from platoons with a lead vehicle piloted by a human driver to fully ad-hoc vehicular networks, a.k.a. VANETs, comprising autonomous/automated vehicles. Safety issues in IVNs appear to be the least studied in the ITS domain. The focus of our work is on safety-critical (SC) scenarios, where accidents and fatalities inevitably occur when such scenarios are not handled correctly. In addition to on-board robotics, inter-vehicular radio communications have been considered for achieving safety properties. Since both technologies have known intrinsic limitations (in addition to possibly experiencing temporary or permanent failures), using them redundantly is mandatory for meeting safety regulations. Redundancy is a fundamental design principle in every SC cyber-physical domain, such as, e.g., air transportation. (Optics-based inter-vehicular communications may also be part of such redundant constructs.) The focus of our on-going work is on safety-critical (SC) communications. We consider IVNs on main roads and highways, which are settings where velocities can be very high, thus exacerbating safety problems acceptable delays in the cyber space, and response times in the physical space, shall be very small. Human lives being at stake, such delays and response times must have strict (non-stochastic) upper bounds under worst-case conditions (vehicular density, concurrency and failures). Consequently, we are led to look for deterministic solutions.

Rationale

In the current ITS literature, the term *safety* is used without being given a precise definition. That must be corrected. In our case, a fundamental open question is: what is the exact meaning of *SC communications*? We have devised a definition, referred to as space-time bounds acceptability (STBA) requirements. For any given problem related to SC communications, those STBA requirements serve as yardsticks for distinguishing acceptable solutions from unacceptable ones with respect to safety. In conformance with the above, STBA requirements rest on the following worst-case upper bounds: λ for channel access delays, and Δ for distributed inter-vehicular coordination (message dissemination, distributed agreement).

Via discussions with foreign colleagues, notably those active in the IEEE 802 Committee, we have comforted our early diagnosis regarding existing standards for V2V/V2I/V2X communications, such as IEEE 802.11p and ETSI ITS-G5: they are totally inappropriate regarding SC communications. A major flaw is the choice of CSMA/CA as the MAC-level protocol. Obviously, there cannot be such bounds as λ and Δ with CSMA/CA. Another flaw is the choice of medium-range omnidirectional communications, radio range in the order of 250 m, and interference range in the order of 400 m. Stochastic delays achievable with existing standards are just unacceptable in moderate/worst-case contention conditions. Consider the following setting, not uncommon in many countries: a highway, 3 lanes each direction, dense traffic, i.e. 1 vehicle per 12.5 m. A simple

calculation leads to the following result: any vehicle may experience (destructive) interferences from up to 384 vehicles. Even if one assumes some reasonable communications activity ratio, say 25%, one finds that up to 96 vehicles may be contending for channel access. Under such conditions, MAC-level delays and string-wide dissemination/agreement delays achieved by current standards fail to meet the STBA requirements by huge margins.

Reliance on V2I communications via terrestrial infrastructures and nodes, such as road-side units or WiFi hotspots, rather than direct V2V communications, can only lead to poorer results. First, reachability is not guaranteed: hazardous conditions may develop anywhere anytime, far away from a terrestrial node. Second, mixing SC communications and ordinary communications within terrestrial nodes is a violation of the very fundamental segregation principle: SC communications and processing shall be isolated from ordinary communications and processing. Third, security: it is very easy to jam or to spy on a terrestrial node; moreover, terrestrial nodes may be used for launching all sorts of attacks, man-in-the-middle attacks for example. Fourth, delays can only get worse than with direct V2V communications, since transiting via a node inevitably introduces additional latencies. Fifth, the delivery of every SC message must be acknowledged, which exacerbates the latency problems. Sixth, availability: what happens when a terrestrial node fails?

Trying to tweak existing standards for achieving SC communications is vain. That is also unjustified. Clearly, medium-range omnidirectional communications are unjustified for the handling of SC scenarios. By definition, accidents can only involve vehicles that are very close to each other. Therefore, short-range directional communications suffice. The obvious conclusion is that novel protocols and inter-vehicular coordination algorithms based on short-range direct V2V communications are needed. It is mandatory to check whether these novel solutions meet the STBA requirements. Future standards specifically aimed at SC communications in IVNs may emerge from such solutions.

Naming and privacy

Additionally, we are exploring the (re)naming problem as it arises in IVNs. Source and destination names appear in messages exchanged among vehicles. Most often, names are IP addresses or MAC addresses (plate numbers shall not be used for privacy reasons). A vehicle which intends to communicate with some vehicle, denoted V here, must know which name *name(V)* to use in order to reach/designate V. Existing solutions are based on multicasting/broadcasting existential messages, whereby every vehicle publicizes its existence (name and geolocation), either upon request (replying to a Geocast) or spontaneously (periodic beaconing). These solutions have severe drawbacks. First, they contribute to overloading communication channels (leading to unacceptably high worst-case delays). Second, they amount to breaching privacy voluntarily. Why should vehicles reveal their existence and their time dependent geolocations, making tracing and spying much easier? Novel solutions are needed. They shall be such that:

- At any time, a vehicle can assign itself a name that is unique within a geographical zone centered on that vehicle (no third-party involved),
- No linkage may exist between a name and those identifiers (plate numbers, IP/MAC addresses, etc.) proper to a vehicle,
- Different (unique) names can be computed at different times by a vehicle (names can be short-lived or long-lived),
- *name(V)* at UTC time t is revealed only to those vehicles sufficiently close to V at time t, notably those which may collide with V.

We have solved the (re)naming problem in string/cohort formations [34]. Ranks (unique integers in any given string/cohort) are privacy-preserving names, easily computed by every member of a string, in the presence of string membership changes (new vehicles join in, members leave). That problem is open when considering arbitrary clusters of vehicles/strings encompassing multiple lanes.

3.3. Probabilistic modeling for large transportation systems

Participants: Guy Fayolle, Jean-Marc Lasgouttes.

This activity concerns the modeling of random systems related to ITS, through the identification and development of solutions based on probabilistic methods and more specifically through the exploration of links between large random systems and statistical physics. Traffic modeling is a very fertile area of application for this approach, both for macroscopic (fleet management [32], traffic prediction) and for microscopic (movement of each vehicle, formation of traffic jams) analysis. When the size or volume of structures grows (leading to the so-called "thermodynamic limit"), we study the quantitative and qualitative (performance, speed, stability, phase transitions, complexity, etc.) features of the system.

In the recent years, several directions have been explored.

3.3.1. Traffic reconstruction

Large random systems are a natural part of macroscopic studies of traffic, where several models from statistical physics can be fruitfully employed. One example is fleet management, where one main issue is to find optimal ways of reallocating unused vehicles: it has been shown that Coulombian potentials might be an efficient tool to drive the flow of vehicles. Another case deals with the prediction of traffic conditions, when the data comes from probe vehicles instead of static sensors.

While the widely-used macroscopic traffic flow models are well adapted to highway traffic, where the distance between junction is long (see for example the work done by the NeCS team in Grenoble), our focus is on a more urban situation, where the graphs are much denser. The approach we are advocating here is model-less, and based on statistical inference rather than fundamental diagrams of road segments. Using the Ising model or even a Gaussian Random Markov Field, together with the very popular Belief Propagation (BP) algorithm, we have been able to show how real-time data can be used for traffic prediction and reconstruction (in the space-time domain).

This new use of BP algorithm raises some theoretical questions about the ways the make the belief propagation algorithm more efficient:

- find the best way to inject real-valued data in an Ising model with binary variables [36];
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information [33];
- make the underlying model as sparse as possible, in order to improve BP convergence and quality [35].

3.3.2. Exclusion processes for road traffic modeling

The focus here is on road traffic modeled as a granular flow, in order to analyze the features that can be explained by its random nature. This approach is complementary to macroscopic models of traffic flow (as done for example in the Opale team at Inria), which rely mainly on ODEs and PDEs to describe the traffic as a fluid.

One particular feature of road traffic that is of interest to us is the spontaneous formation of traffic jams. It is known that systems as simple as the Nagel-Schreckenberg model are able to describe traffic jams as an emergent phenomenon due to interaction between vehicles. However, even this simple model cannot be explicitly analyzed and therefore one has to resort to simulation.

One of the simplest solvable (but non trivial) probabilistic models for road traffic is the exclusion process. It lends itself to a number of extensions allowing to tackle some particular features of traffic flows: variable speed of particles, synchronized move of consecutive particles (platooning), use of geometries more complex than plain 1D (cross roads or even fully connected networks), formation and stability of vehicle clusters (vehicles that are close enough to establish an ad-hoc communication system), two-lane roads with overtaking.

The aspect that we have particularly studied is the possibility to let the speed of vehicle evolve with time. To this end, we consider models equivalent to a series of queues where the pair (service rate, number of customers) forms a random walk in the quarter plane $\mathbb{Z}_+^2$.

Having in mind a global project concerning the analysis of complex systems, we also focus on the interplay between discrete and continuous description: in some cases, this recurrent question can be addressed quite rigorously via probabilistic methods.

We have considered in [30] some classes of models dealing with the dynamics of discrete curves subjected to stochastic deformations. It turns out that the problems of interest can be set in terms of interacting exclusion processes, the ultimate goal being to derive hydrodynamic limits after proper scaling. A seemingly new method is proposed, which relies on the analysis of specific partial differential operators, involving variational calculus and functional integration. Starting from a detailed analysis of the Asymmetric Simple Exclusion Process (ASEP) system on the torus $\mathbb{Z}/n\mathbb{Z}$, the arguments a priori work in higher dimensions (ABC, multi-type exclusion processes, etc), leading to systems of coupled partial differential equations of Burgers' type.

3.3.3. Random walks in the quarter plane $\mathbb{Z}_+^2$

This field remains one of the important *violon d'Ingres* in our research activities in stochastic processes, both from theoretical and applied points of view. In particular, it is a building block for models of many communication and transportation systems.

One essential question concerns the computation of stationary measures (when they exist). As for the answer, it has been given by original methods formerly developed in the team (see books and related bibliography). For instance, in the case of small steps (jumps of size one in the interior of $\mathbb{Z}_+^2$), the invariant measure $\{\pi_{i,j}, i,j \geq 0\}$ does satisfy the fundamental functional equation (see [2]):

$$Q(x,y)\pi(x,y) = q(x,y)\pi(x) + \widetilde{q}(x,y)\widetilde{\pi}(y) + \pi_0(x,y). \tag{68}$$

where the unknown generating functions $\pi(x,y), \pi(x), \widetilde{\pi}(y), \pi_0(x,y)$ are sought to be analytic in the region $\{(x,y) \in \mathbb{C}^2 : |x| < 1, |y| < 1\}$, and continuous on their respective boundaries.

The given function $Q(x,y) = \sum_{i,j} p_{i,j} x^i y^j - 1$, where the sum runs over the possible jumps of the walk inside $\mathbb{Z}_+^2$, is often referred to as the *kernel*. Then it has been shown that equation (1) can be solved by reduction to a boundary-value problem of Riemann-Hilbert type. This method has been the source of numerous and fruitful developments. Some recent and ongoing works have been dealing with the following matters.

- *Group of the random walk*. In several studies, it has been noticed that the so-called *group of the walk* governs the behavior of a number of quantities, in particular through its *order*, which is always even. In the case of small jumps, the algebraic curve R defined by $\{Q(x,y) = 0\}$ is either of *genus* 0 (the sphere) or 1 (the torus). In [Fayolle-2011a], when the drift of the random walk is equal to 0 (and then so is the genus), an effective criterion gives the *order* of the group. More generally, it is also proved that whenever the genus is 0, this order is infinite, except precisely for the zero drift case, where finiteness is quite possible. When the *genus* is 1, the situation is more difficult. Recently [31], a criterion has been found in terms of a determinant of order 3 or 4, depending on the arity of the group.
- *Nature of the counting generating functions*. Enumeration of planar lattice walks is a classical topic in combinatorics. For a given set of allowed jumps (or steps), it is a matter of counting the number of paths starting from some point and ending at some arbitrary point in a given time, and possibly restricted to some regions of the plane. A first basic and natural question arises: how many such paths exist? A second question concerns the nature of the associated counting generating functions (CGF): are they rational, algebraic, holonomic (or D-finite, i.e. solution of a linear differential equation with polynomial coefficients)?

  Let $f(i,j,k)$ denote the number of paths in $\mathbb{Z}_+^2$ starting from $(0,0)$ and ending at (i,j) at time k. Then the corresponding CGF

$$F(x,y,z) = \sum_{i,j,k \geq 0} f(i,j,k) x^i y^j z^k \tag{69}$$

satisfies the functional equation

$$K(x,y)F(x,y,z) = c(x)F(x,0,z) + \widetilde{c}(y)F(0,y,z) + c_0(x,y), \tag{70}$$

where z is considered as a time-parameter. Clearly, equations (2) and (1) are of the same nature, and answers to the above questions have been given in [Fayolle-2010].

- *Some exact asymptotics in the counting of walks in* $\mathbb{Z}_+^2$. A new and uniform approach has been proposed about the following problem: *What is the asymptotic behavior, as their length goes to infinity, of the number of walks ending at some given point or domain (for instance one axis)?* The method in [Fayolle-2012] works for *both* finite or infinite groups, and for walks not necessarily restricted to excursions.

3.3.4. Simulation for urban mobility

We have worked on various simulation tools to study and evaluate the performance of different transportation modes covering an entire urban area.

- Discrete event simulation for collective taxis, a public transportation system with a service quality comparable with that of conventional taxis.
- Discrete event simulation a system of self-service cars that can reconfigure themselves into shuttles, therefore creating a multimodal public transportation system; this second simulator is intended to become a generic tool for multimodal transportation.
- Joint microscopic simulation of mobility and communication, necessary for investigation of cooperative platoons performance.

These two programs use a technique allowing to run simulations in batch mode and analyze the dynamics of the system afterward.

SEMAGRAMME Project-Team

3. Research Program

3.1. Overview

The research program of Sémagramme aims to develop models based on well-established mathematics. We seek two main advantages from this approach. On the one hand, by relying on mature theories, we have at our disposal sets of mathematical tools that we can use to study our models. On the other hand, developing various models on a common mathematical background will make them easier to integrate, and will ease the search for unifying principles.

The main mathematical domains on which we rely are formal language theory, symbolic logic, and type theory.

3.2. Formal Language Theory

Formal language theory studies the purely syntactic and combinatorial aspects of languages, seen as sets of strings (or possibly trees or graphs). Formal language theory has been especially fruitful for the development of parsing algorithms for context-free languages. We use it, in a similar way, to develop parsing algorithms for formalisms that go beyond context-freeness. Language theory also appears to be very useful in formally studying the expressive power and the complexity of the models we develop.

3.3. Symbolic Logic

Symbolic logic (and, more particularly, proof-theory) is concerned with the study of the expressive and deductive power of formal systems. In a rule-based approach to computational linguistics, the use of symbolic logic is ubiquitous. As we previously said, at the level of syntax, several kinds of grammars (generative, categorial...) may be seen as basic deductive systems. At the level of semantics, the meaning of an utterance is captured by computing (intermediate) semantic representations that are expressed as logical forms. Finally, using symbolic logics allows one to formalize notions of inference and entailment that are needed at the level of pragmatics.

3.4. Type Theory and Typed λ-Calculus

Among the various possible logics that may be used, Church's simply typed λ-calculus and simple theory of types (a.k.a. higher-order logic) play a central part. On the one hand, Montague semantics is based on the simply typed λ-calculus, and so is our syntax-semantics interface model. On the other hand, as shown by Gallin, the target logic used by Montague for expressing meanings (i.e., his intensional logic) is essentially a variant of higher-order logic featuring three atomic types (the third atomic type standing for the set of possible worlds).

SIROCCO Project-Team

3. Research Program

3.1. Introduction

The research activities on analysis, compression and communication of visual data mostly rely on tools and formalisms from the areas of statistical image modeling, of signal processing, of machine learning, of coding and information theory. Some of the proposed research axes are also based on scientific foundations of computer vision (e.g. multi-view modeling and coding). We have limited this section to some tools which are central to the proposed research axes, but the design of complete compression and communication solutions obviously rely on a large number of other results in the areas of motion analysis, transform design, entropy code design, etc which cannot be all described here.

3.2. Data Dimensionality Reduction

Manifolds, graph-based transforms, compressive sensing

Dimensionality reduction encompasses a variety of methods for low-dimensional data embedding, such as sparse and low-rank models, random low-dimensional projections in a compressive sensing framework, and sparsifying transforms including graph-based transforms. These methods are the cornerstones of many visual data processing tasks (compression, inverse problems).

Sparse representations, *compressive sensing*, and *dictionary learning* have been shown to be powerful tools for efficient processing of visual data. The objective of *sparse representations* is to find a sparse approximation of a given input data. In theory, given a dictionary matrix $A \in \mathbb{R}^{m \times n}$, and a data $\mathbf{b} \in \mathbb{R}^m$ with $m << n$ and A is of full row rank, one seeks the solution of $\min\{\|\mathbf{x}\|_0 \ : \ A\mathbf{x} = \mathbf{b}\}$, where $\|\mathbf{x}\|_0$ denotes the ℓ_0 norm of $\mathbf{x}$, i.e. the number of non-zero components in $\mathbf{x}$. A is known as the dictionary, its columns a_j are the atoms, they are assumed to be normalized in Euclidean norm. There exist many solutions x to $Ax = b$. The problem is to find the sparsest solution x, i.e. the one having the fewest nonzero components. In practice, one actually seeks an approximate and thus even sparser solution which satisfies $\min\{\|\mathbf{x}\|_0 \ : \ \|A\mathbf{x} - \mathbf{b}\|_p \leq \rho\}$, for some $\rho \geq 0$, characterizing an admissible reconstruction error.

The recent theory of *compressed sensing*, in the context of discrete signals, can be seen as an effective dimensionality reduction technique. The idea behind compressive sensing is that a signal can be accurately recovered from a small number of linear measurements, at a rate much smaller than what is commonly prescribed by the Shannon-Nyquist theorem, provided that it is sparse or compressible in a known basis. Compressed sensing has emerged as a powerful framework for signal acquisition and sensor design, with a number of open issues such as learning the basis in which the signal is sparse, with the help of dictionary learning methods, or the design and optimization of the sensing matrix. The problem is in particular investigated in the context of light fields acquisition, aiming at novel camera design with the goal of offering a good trade-off between spatial and angular resolution.

While most image and video processing methods have been developed for cartesian sampling grids, new imaging modalities (e.g. point clouds, light fields) call for representations on irregular supports that can be well represented by *graphs*. Reducing the dimensionality of such signals require designing novel transforms yielding compact signal representation. One example of transform is the Graph Fourier transform whose basis functions are given by the eigenvectors of the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal degree matrix whose i^{th} diagonal element is equal to the sum of the weights of all edges incident to the node i, and $\mathbf{A}$ the adjacency matrix. The eigenvectors of the Laplacian of the graph, also called Laplacian eigenbases, are analogous to the Fourier bases in the Euclidean domain and allow representing the signal residing on the graph as a linear combination of eigenfunctions akin to Fourier Analysis. This transform is particularly efficient for compacting smooth signals on the graph. The problems which therefore need to be addressed are (i) to define graph structures on which the corresponding signals are smooth for different imaging modalities and (ii) the design of transforms compacting well the signal energy with a tractable computational complexity.

3.3. Deep neural networks

Autoencoders, Neural Networks, Recurrent Neural Networks

From dictionary learning which we have investigated a lot in the past, our activity is now evolving towards deep learning techniques which we are considering for dimensionality reduction. We address the problem of unsupervised learning of transforms and prediction operators that would be optimal in terms of energy compaction, considering autoencoders and neural network architectures.

An autoencoder is a neural network with an encoder g_e, parametrized by θ, that computes a representation Y from the data X, and a decoder g_d, parametrized by ϕ, that gives a reconstruction $\widehat{X}$ of X (see Figure below). Autoencoders can be used for dimensionality reduction, compression, denoising. When it is used for compression, the representation need to be quantized, leading to a quantized representation $\widehat{Y} = Q(Y)$ (see Figure below). If an autoencoder has fully-connected layers, the architecture, and the number of parameters to be learned, depends on the image size. Hence one autoencoder has to be trained per image size, which poses problems in terms of genericity.

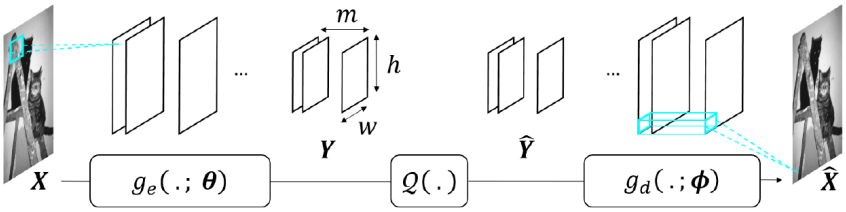


Figure 1. Illustration of an autoencoder.

To avoid this limitation, architectures without fully-connected layer and comprising instead convolutional layers and non-linear operators, forming convolutional neural networks (CNN) may be preferrable. The obtained representation is thus a set of so-called feature maps.

The other problems that we address with the help of neural networks are scene geometry and scene flow estimation, view synthesis, prediction and interpolation with various imaging modalities. The problems are posed either as supervised or unsupervised learning tasks. Our scope of investigation includes autoencoders, convolutional networks, variational autoencoders and generative adversarial networks (GAN) but also recurrent networks and in particular Long Short Term Memory (LSTM) networks. Recurrent neural networks attempting to model time or sequence dependent behaviour, by feeding back the output of a neural network layer at time t to the input of the same network layer at time t+1, have been shown to be interesting tools for temporal frame prediction. LSTMs are particular cases of recurrent networks made of cells composed of three types of neural layers called gates.

Deep neural networks have also been shown to be very promising for solving inverse problems (e.g. super-resolution, sparse recovery in a compressive sensing framework, inpainting) in image processing. Variational autoencoders, generative adversarial networks (GAN), learn, from a set of examples, the latent space or the manifold in which the images, that we search to recover, reside. The inverse problems can be re-formulated using a regularization in the latent space learned by the network. For the needs of the regularization, the learned latent space may need to verify certain properties such as preserving distances or neighborhood of the input space, or in terms of statistical modeling. GANs, trained to produce images that are plausible, are also useful tools for learning texture models, expressed via the filters of the network, that can be used for solving problems like inpainting or view synthesis.

3.4. Coding theory

OPTA limit (Optimum Performance Theoretically Attainable), Rate allocation, Rate-Distortion optimization, lossy coding, joint source-channel coding multiple description coding, channel modelization, oversampled frame expansions, error correcting codes.

Source coding and channel coding theory [0] is central to our compression and communication activities, in particular to the design of entropy codes and of error correcting codes. Another field in coding theory which has emerged in the context of sensor networks is Distributed Source Coding (DSC). It refers to the compression of correlated signals captured by different sensors which do not communicate between themselves. All the signals captured are compressed independently and transmitted to a central base station which has the capability to decode them jointly. DSC finds its foundation in the seminal Slepian-Wolf [0] (SW) and Wyner-Ziv [0] (WZ) theorems. Let us consider two binary correlated sources X and Y. If the two coders communicate, it is well known from Shannon's theory that the minimum lossless rate for X and Y is given by the joint entropy $H(X,Y)$. Slepian and Wolf have established in 1973 that this lossless compression rate bound can be approached with a vanishing error probability for long sequences, even if the two sources are coded separately, provided that they are decoded jointly and that their correlation is known to both the encoder and the decoder.

In 1976, Wyner and Ziv considered the problem of coding of two correlated sources X and Y, with respect to a fidelity criterion. They have established the rate-distortion function $R*_{X|Y}(D)$ for the case where the side information Y is perfectly known to the decoder only. For a given target distortion D, $R*_{X|Y}(D)$ in general verifies $R_{X|Y}(D) \leq R*_{X|Y}(D) \leq R_X(D)$, where $R_{X|Y}(D)$ is the rate required to encode X if Y is available to both the encoder and the decoder, and R_X is the minimal rate for encoding X without SI. These results give achievable rate bounds, however the design of codes and practical solutions for compression and communication applications remain a widely open issue.

[0] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, July 2006.

[0] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources." IEEE Transactions on Information Theory, 19(4), pp. 471-480, July 1973.

[0] A. Wyner and J. Ziv, "The rate-distortion function for source coding ith side information at the decoder." IEEE Transactions on Information Theory, pp. 1-10, January 1976.

Stars Project-Team

3. Research Program

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, action recognition and semantic activity recognition. **These three research directions are organized following the workflow of activity recognition systems**: First, *the perception* and *the action recognition* directions provide new techniques to extract powerful features, whereas *the semantic activity recognition* research direction provides new paradigms to match these features with concrete video analytic and healthcare applications.

Transversely, we consider a *new research axis in machine learning*, combining a priori knowledge and learning techniques, to set up the various models of an activity recognition system. A major objective is to automate model building or model enrichment at the perception level and at the understanding level.

3.2. Perception for Activity Recognition

Participants: François Brémond, Antitza Dantcheva, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Machine Learning, Computer Vision, Cognitive Vision Systems, Software Engineering

3.2.1. Introduction

Our main goal in perception is to develop vision algorithms able to address the large variety of conditions characterizing real world scenes in terms of sensor conditions, hardware requirements, lighting conditions, physical objects, and application objectives. We have also several issues related to perception which combine machine learning and perception techniques: learning people appearance, parameters for system control and shape statistics.

3.2.2. Appearance Models and People Tracking

An important issue is to detect in real-time physical objects from perceptual features and predefined 3D models. It requires finding a good balance between efficient methods and precise spatio-temporal models. Many improvements and analysis need to be performed in order to tackle the large range of people detection scenarios.

Appearance models. In particular, we study the temporal variation of the features characterizing the appearance of a human. This task could be achieved by clustering potential candidates depending on their position and their reliability. This task can provide any people tracking algorithms with reliable features allowing for instance to (1) better track people or their body parts during occlusion, or to (2) model people appearance for re-identification purposes in mono and multi-camera networks, which is still an open issue. The underlying challenge of the person re-identification problem arises from significant differences in illumination, pose and camera parameters. The re-identification approaches have two aspects: (1) establishing correspondences between body parts and (2) generating signatures that are invariant to different color responses. As we have already several descriptors which are color invariant, we now focus more on aligning two people detection and on finding their corresponding body parts. Having detected body parts, the approach can handle pose variations. Further, different body parts might have different influence on finding the correct match among a whole gallery dataset. Thus, the re-identification approaches have to search for matching strategies. As the results of the re-identification are always given as the ranking list, re-identification focuses on learning to rank. "Learning to rank" is a type of machine learning problem, in which the goal is to automatically construct a ranking model from a training data.

Therefore, we work on information fusion to handle perceptual features coming from various sensors (several cameras covering a large scale area or heterogeneous sensors capturing more or less precise and rich information). New 3D RGB-D sensors are also investigated, to help in getting an accurate segmentation for specific scene conditions.

Long term tracking. For activity recognition we need robust and coherent object tracking over long periods of time (often several hours in video surveillance and several days in healthcare). To guarantee the long term coherence of tracked objects, spatio-temporal reasoning is required. Modeling and managing the uncertainty of these processes is also an open issue. In Stars we propose to add a reasoning layer to a classical Bayesian framework modeling the uncertainty of the tracked objects. This reasoning layer can take into account the a priori knowledge of the scene for outlier elimination and long-term coherency checking.

Controlling system parameters. Another research direction is to manage a library of video processing programs. We are building a perception library by selecting robust algorithms for feature extraction, by insuring they work efficiently with real time constraints and by formalizing their conditions of use within a program supervision model. In the case of video cameras, at least two problems are still open: robust image segmentation and meaningful feature extraction. For these issues, we are developing new learning techniques.

3.3. Action Recognition

Participants: François Brémond, Antitza Dantcheva, Monique Thonnat.

Machine Learning, Computer Vision, Cognitive Vision Systems

3.3.1. Introduction

Due to the recent development of high processing units, such as GPU, this is now possible to extract meaningful features directly from videos (e.g. video volume) to recognize reliably short actions. Action Recognition benefits also greatly from the huge progress made recently in Machine Learning (e.g. Deep Learning), especially for the study of human behavior. For instance, Action Recognition enables to measure objectively the behavior of humans by extracting powerful features characterizing their everyday activities, their emotion, eating habits and lifestyle, by learning models from a large number of data from a variety of sensors, to improve and optimize for example, the quality of life of people suffering from behavior disorders. However, Smart Homes and Partner Robots have been well advertised but remain laboratory prototypes, due to the poor capability of automated systems to perceive and reason about their environment. A hard problem is for an automated system to cope 24/7 with the variety and complexity of the real world. Another challenge is to extract people fine gestures and subtle facial expressions to better analyze behavior disorders, such as anxiety or apathy. Taking advantage of what is currently studied for self-driving cars or smart retails, there is a large avenue to design ambitious approaches for the healthcare domain. In particular, the advance made with Deep Learning algorithms has already enabled to recognize complex activities, such as cooking interactions with instruments, and from this analysis to differentiate healthy people from the ones suffering from dementia.

To address these issues, we propose to tackle several challenges:

3.3.2. Action recognition in the wild

The current Deep Learning techniques are mostly developed to work on few clipped videos, which have been recorded with students performing a limited set of predefined actions in front of a camera with high resolution. However, real life scenarios include actions performed in a spontaneous manner by older people (including people interactions with their environment or with other people), from different viewpoints, with varying framerate, partially occluded by furniture at different locations within an apartment depicted through long untrimmed videos. Therefore, a new dedicated dataset should be collected in a real-world setting to become a public benchmark video dataset and to design novel algorithms for ADL activity recognition. A special attention should be taken to anonymize the videos.

3.3.3. Attention mechanisms for action recognition

Activities of Daily Living (ADL) and video-surveillance activities are different from internet activities (e.g. Sports, Movies, YouTube), as they may have very similar context (e.g. same background kitchen) with high intra-variation (different people performing the same action in different manners), but in the same time low inter-variation, similar ways to perform two different actions (e.g. eating and drinking a glass of water). Consequently, fine-grained actions are badly recognized. So, we will design novel attention mechanisms for action recognition, for the algorithm being able to focus on a discriminative part of the person conducting the action. For instance, we will study attention algorithms, which could focus on the most appropriate body parts (e.g. full body, right hand). In particular, we plan to design a soft mechanism, learning the attention weights directly on the feature map of a 3DconvNet, a powerful convolutional network, which takes as input a batch of videos.

3.3.4. Action detection for untrimmed videos

Many approaches have been proposed to solve the problem of action recognition in short clipped 2D videos, which achieved impressive results with hand-crafted and deep features. However, these approaches cannot address real life situations, where cameras provide online and continuous video streams in applications such as robotics, video surveillance, and smart-homes. Here comes the importance of action detection to help recognizing and localizing each action happening in long videos. Action detection can be defined as the ability to localize starting and ending of each human action happening in the video, in addition to recognizing each action label. There have been few action detection algorithms designed for untrimmed videos, which are based on either sliding window, temporal pooling or frame-based labeling. However, their performance is too low to address real-word datasets. A first task consists in benchmarking the already published approaches to study their limitations on novel untrimmed video datasets, recorded following real-world settings. A second task could be to propose a new mechanism to improve either 1) the temporal pooling directly from the 3DconvNet architecture using for instance Temporal Convolution Networks (TCNs) or 2) frame-based labeling with a clustering technique (e.g. using Fisher Vectors) to discover the sub-activities of interest.

3.3.5. View invariant action recognition

The performance of current approaches strongly relies on the used camera angle: enforcing that the camera angle used in testing is the same (or extremely close to) as the camera angle used in training, is necessary for the approach performs well. On the contrary, the performance drops when a different camera view-point is used. Therefore, we aim at improving the performance of action recognition algorithms by relying on 3D human pose information. For the extraction of the 3D pose information, several open-source algorithms can be used, such as openpose or videopose3D (from CMU or Facebook research, https://github.com/CMU-Perceptual-Computing-Lab/openpose. Also, other algorithms extracting 3d meshes can be used. To generate extra views, Generative Adversial Network (GAN) can be used together with the 3D human pose information to complete the training dataset from the missing view.

3.3.6. Uncertainty and action recognition

Another challenge is to combine the short-term actions recognized by powerful Deep Learning techniques with long-term activities defined by constraint-based descriptions and linked to user interest. To realize this objective, we have to compute the uncertainty (i.e. likelihood or confidence), with which the short-term actions are inferred. This research direction is linked to the next one, to Semantic Activity Recognition.

3.4. Semantic Activity Recognition

Participants: François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Computer Vision

3.4.1. Introduction

Semantic activity recognition is a complex process where information is abstracted through four levels: signal (e.g. pixel, sound), perceptual features, physical objects and activities. The signal and the feature levels are characterized by strong noise, ambiguous, corrupted and missing data. The whole process of scene understanding consists in analyzing this information to bring forth pertinent insight of the scene and its dynamics while handling the low level noise. Moreover, to obtain a semantic abstraction, building activity models is a crucial point. A still open issue consists in determining whether these models should be given a priori or learned. Another challenge consists in organizing this knowledge in order to capitalize experience, share it with others and update it along with experimentation. To face this challenge, tools in knowledge engineering such as machine learning or ontology are needed.

Thus we work along the following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models), learning (how to learn the models needed for activity recognition) and activity recognition and discrete event systems.

3.4.2. High Level Understanding

A challenging research axis is to recognize subjective activities of physical objects (i.e. human beings, animals, vehicles) based on a priori models and objective perceptual measures (e.g. robust and coherent object tracks).

To reach this goal, we have defined original activity recognition algorithms and activity models. Activity recognition algorithms include the computation of spatio-temporal relationships between physical objects. All the possible relationships may correspond to activities of interest and all have to be explored in an efficient way. The variety of these activities, generally called video events, is huge and depends on their spatial and temporal granularity, on the number of physical objects involved in the events, and on the event complexity (number of components constituting the event).

Concerning the modeling of activities, we are working towards two directions: the uncertainty management for representing probability distributions and knowledge acquisition facilities based on ontological engineering techniques. For the first direction, we are investigating classical statistical techniques and logical approaches. For the second direction, we built a language for video event modeling and a visual concept ontology (including color, texture and spatial concepts) to be extended with temporal concepts (motion, trajectories, events ...) and other perceptual concepts (physiological sensor concepts ...).

3.4.3. Learning for Activity Recognition

Given the difficulty of building an activity recognition system with a priori knowledge for a new application, we study how machine learning techniques can automate building or completing models at the perception level and at the understanding level.

At the understanding level, we are learning primitive event detectors. This can be done for example by learning visual concept detectors using SVMs (Support Vector Machines) with perceptual feature samples. An open question is how far can we go in weakly supervised learning for each type of perceptual concept (i.e. leveraging the human annotation task). A second direction is to learn typical composite event models for frequent activities using trajectory clustering or data mining techniques. We name composite event a particular combination of several primitive events.

3.4.4. Activity Recognition and Discrete Event Systems

The previous research axes are unavoidable to cope with the semantic interpretations. However they tend to let aside the pure event driven aspects of scenario recognition. These aspects have been studied for a long time at a theoretical level and led to methods and tools that may bring extra value to activity recognition, the most important being the possibility of formal analysis, verification and validation.

We have thus started to specify a formal model to define, analyze, simulate, and prove scenarios. This model deals with both absolute time (to be realistic and efficient in the analysis phase) and logical time (to benefit from well-known mathematical models providing re-usability, easy extension, and verification). Our purpose is to offer a generic tool to express and recognize activities associated with a concrete language to specify activities in the form of a set of scenarios with temporal constraints. The theoretical foundations and the tools being shared with Software Engineering aspects.

The results of the research performed in perception and semantic activity recognition (first and second research directions) produce new techniques for scene understanding and contribute to specify the needs for new software architectures (third research direction).

THOTH Project-Team

3. Research Program

3.1. Designing and learning structured models

The task of understanding image and video content has been interpreted in several ways over the past few decades, namely image classification, detecting objects in a scene, recognizing objects and their spatial extents in an image, estimating human poses, recovering scene geometry, recognizing activities performed by humans. However, addressing all these problems individually provides us with a partial understanding of the scene at best, leaving much of the visual data unexplained.

One of the main goals of this research axis is to go beyond the initial attempts that consider only a subset of tasks jointly, by developing novel models for a more complete understanding of scenes to address all the component tasks. We propose to incorporate the structure in image and video data explicitly into the models. In other words, our models aim to satisfy the complex sets of constraints that exist in natural images and videos. Examples of such constraints include: (i) relations between objects, like signs for shops indicate the presence of buildings, people on a road are usually walking or standing, (ii) higher-level semantic relations involving the type of scene, geographic location, and the plausible actions as a global constraint, e.g., an image taken at a swimming pool is unlikely to contain cars, (iii) relating objects occluded in some of the video frames to content in other frames, where they are more clearly visible as the camera or the object itself move, with the use of long-term trajectories and video object proposals.

This research axis will focus on three topics. The first is developing deep features for video. This involves designing rich features available in the form of long-range temporal interactions among pixels in a video sequence to learn a representation that is truly spatio-temporal in nature. The focus of the second topic is the challenging problem of modeling human activities in video, starting from human activity descriptors to building intermediate spatio-temporal representations of videos, and then learning the interactions among humans, objects and scenes temporally. The last topic is aimed at learning models that capture the relationships among several objects and regions in a single image scene, and additionally, among scenes in the case of an image collection or a video. The main scientific challenges in this topic stem from learning the structure of the probabilistic graphical model as well as the parameters of the cost functions quantifying the relationships among its entities. In the following we will present work related to all these three topics and then elaborate on our research directions.

- **Deep features for vision.** Deep learning models provide a rich representation of complex objects but in return have a large number of parameters. Thus, to work well on difficult tasks, a large amount of data is required. In this context, video presents several advantages: objects are observed from a large range of viewpoints, motion information allows the extraction of moving objects and parts, and objects can be differentiated by their motion patterns. We initially plan to develop deep features for videos that incorporate temporal information at multiple scales. We then plan to further exploit the rich content in video by incorporating additional cues, such as the detection of people and their body-joint locations in video, minimal prior knowledge of the object of interest, with the goal of learning a representation that is more appropriate for video understanding. In other words, a representation that is learned from video data and targeted at specific applications. For the application of recognizing human activities, this involves learning deep features for humans and their body-parts with all their spatiotemporal variations, either directly from raw video data or “pre-processed” videos containing human detections. For the application of object tracking, this task amounts to learning object-specific deep representations, further exploiting the limited annotation provided to identify the object.
- **Modeling human activities in videos.** Humans and their activities are not only one of the most frequent and interesting subjects in videos but also one of the hardest to analyze owing to the

complexity of the human form, clothing and movements. As part of this task, the Thoth project-team plans to build on state-of-the-art approaches for spatio-temporal representation of videos. This will involve using the dominant motion in the scene as well as the local motion of individual parts undergoing a rigid motion. Such motion information also helps in reasoning occlusion relationships among people and objects, and the state of the object. This novel spatio-temporal representation ultimately provides the equivalent of object proposals for videos, and is an important component for learning algorithms using minimal supervision. To take this representation even further, we aim to integrate the proposals and the occlusion relationships with methods for estimating human pose in videos, thus leveraging the interplay among body-joint locations, objects in the scene, and the activity being performed. For example, the locations of shoulder, elbow and wrist of a person drinking coffee are constrained to move in a certain way, which is completely different from the movement observed when a person is typing. In essence, this step will model human activities by dynamics in terms of both low-level movements of body-joint locations and global high-level motion in the scene.

- **Structured models.** The interactions among various elements in a scene, such as, the objects and regions in it, the motion of object parts or entire objects themselves, form a key element for understanding image or video content. These rich cues define the structure of visual data and how it evolves spatio-temporally. We plan to develop a novel graphical model to exploit this structure. The main components in this graphical model are spatio-temporal regions (in the case of video or simply image regions), which can represent object parts or entire objects themselves, and the interactions among several entities. The dependencies among the scene entities are defined with a higher order or a global cost function. A higher order constraint is a generalization of the pairwise interaction term, and is a cost function involving more than two components in the scene, e.g., several regions, whereas a global constraint imposes a cost term over the entire image or video, e.g., a prior on the number of people expected in the scene. The constraints we plan to include generalize several existing methods, which are limited to pairwise interactions or a small restrictive set of higher-order costs. In addition to learning the parameters of these novel functions, we will focus on learning the structure of the graph itself—a challenging problem that is seldom addressed in current approaches. This provides an elegant way to go beyond state-of-the-art deep learning methods, which are limited to learning the high-level interaction among parts of an object, by learning the relationships among objects.

3.2. Learning of visual models from minimal supervision

Today's approaches to visual recognition learn models for a limited and fixed set of visual categories with fully supervised classification techniques. This paradigm has been adopted in the early 2000's, and within it enormous progress has been made over the last decade.

The scale and diversity in today's large and growing image and video collections (such as, e.g., broadcast archives, and personal image/video collections) call for a departure from the current paradigm. This is the case because to answer queries about such data, it is unfeasible to learn the models of visual content by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions. For one, it will be difficult, or even impossible to decide a-priori what are the relevant categories and the proper granularity level. Moreover, the cost of such annotations would be prohibitive in most application scenarios. One of the main goals of the Thoth project-team is to develop a new framework for learning visual recognition models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content), and exploiting the weak supervisory signal provided by the accompanying metadata (such as captions, keywords, tags, subtitles, or scripts) and audio signal (from which we can for example extract speech transcripts, or exploit speaker recognition models).

Textual metadata has traditionally been used to index and search for visual content. The information in metadata is, however, typically sparse (e.g., the location and overall topic of newscasts in a video archive [0])

[0]For example at the Dutch national broadcast archive Netherlands Institute of Sound and Vision, with whom we collaborated in the EU FP7 project AXES, typically one or two sentences are used in the metadata to describe a one hour long TV program.

and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off screen or not have survived the final cut). For this reason, metadata search should be complemented by visual content based search, where visual recognition models are used to localize content of interest that is not mentioned in the metadata, to increase the usability and value of image/video archives. *The key insight that we build on in this research axis is that while the metadata for a single image or video is too sparse and noisy to rely on for search, the metadata associated with large video and image databases collectively provide an extremely versatile source of information to learn visual recognition models.* This form of "embedded annotation" is rich, diverse and abundantly available. Mining these correspondences from the web, TV and film archives, and online consumer generated content sites such as Flickr, Facebook, or YouTube, guarantees that the learned models are representative for many different situations, unlike models learned from manually collected fully supervised training data sets which are often biased.

The approach we propose to address the limitations of the fully supervised learning paradigm aligns with "Big Data" approaches developed in other areas: we rely on the orders-of-magnitude-larger training sets that have recently become available with metadata to compensate for less explicit forms of supervision. This will form a sustainable approach to learn visual recognition models for a much larger set of categories with little or no manual intervention. Reducing and ultimately removing the dependency on manual annotations will dramatically reduce the cost of learning visual recognition models. This in turn will allow such models to be used in many more applications, and enable new applications based on visual recognition beyond a fixed set of categories, such as natural language based querying for visual content. This is an ambitious goal, given the sheer volume and intrinsic variability of the every day visual content available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities.

This research axis is organized into the following three sub-tasks:

- **Weakly supervised learning.** For object localization we will go beyond current methods that learn one category model at a time and develop methods that learn models for different categories concurrently. This allows "explaining away" effects to be leveraged, i.e., if a certain region in an image has been identified as an instance of one category, it cannot be an instance of another category at the same time. For weakly supervised detection in video we will consider detection proposal methods. While these are effective for still images, recent approaches for the spatio-temporal domain need further improvements to be similarly effective. Furthermore, we will exploit appearance and motion information jointly over a set of videos. In the video domain we will also continue to work on learning recognition models from subtitle and script information. The basis of leveraging the script data which does not have a temporal alignment with the video, is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited amount of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.
- **Online learning of visual models.** As a larger number of visual category models is being learned, online learning methods become important, since new training data and categories will arrive over time. We will develop online learning methods that can incorporate new examples for existing category models, and learn new category models from few examples by leveraging similarity to related categories using multi-task learning methods. Here we will develop new distance-based classifiers and attribute and label embedding techniques, and explore the use of NLP techniques such as skipgram models to automatically determine between which classes transfer should occur. Moreover, NLP will be useful in the context of learning models for many categories to identify synonyms, and to determine cases of polysemy (e.g. jaguar car brand v.s. jaguar animal), and merge or refine categories accordingly. Ultimately this will result in methods that are able to learn an"encyclopedia" of visual models.
- **Visual search from unstructured textual queries.** We will build on recent approaches that learn

recognition models on-the-fly (as the query is issued) from generic image search engines such as Google Images. While it is feasible to learn models in this manner in a matter of seconds, it is challenging to use the model to retrieve relevant content in real-time from large video archives of more than a few thousand hours. To achieve this requires feature compression techniques to store visual representations in memory, and cascaded search techniques to avoid exhaustive search. This approach, however, leaves untouched the core problem of how to associate visual material with the textual query in the first place. The second approach we will explore is based on image annotation models. In particular we will go beyond image-text retrieval methods by using recurrent neural networks such as Elman networks or long short-term memory (LSTM) networks to generate natural language sentences to describe images.

3.3. Large-scale learning and optimization

We have entered an era of massive data acquisition, leading to the revival of an old scientific utopia: it should be possible to better understand the world by automatically converting data into knowledge. It is also leading to a new economic paradigm, where data is a valuable asset and a source of activity. Therefore, developing scalable technology to make sense of massive data has become a strategic issue. Computer vision has already started to adapt to these changes.

In particular, very high dimensional models such as deep networks are becoming highly popular and successful for visual recognition. This change is closely related to the advent of big data. On the one hand, these models involve a huge number of parameters and are rich enough to represent well complex objects such as natural images or text corpora. On the other hand, they are prone to overfitting (fitting too closely to training data without being able to generalize to new unseen data) despite regularization; to work well on difficult tasks, they require a large amount of labelled data that has been available only recently. Other cues may explain their success: the deep learning community has made significant engineering efforts, making it possible to learn in a day on a GPU large models that would have required weeks of computations on a traditional CPU, and it has accumulated enough empirical experience to find good hyper-parameters for its networks.

To learn the huge number of parameters of deep hierarchical models requires scalable optimization techniques and large amounts of data to prevent overfitting. This immediately raises two major challenges: how to learn without large amounts of labeled data, or with weakly supervised annotations? How to efficiently learn such huge-dimensional models? To answer the above challenges, we will concentrate on the design and theoretical justifications of deep architectures including our recently proposed deep kernel machines, with a focus on weakly supervised and unsupervised learning, and develop continuous and discrete optimization techniques that push the state of the art in terms of speed and scalability.

This research axis will be developed into three sub-tasks:

- **Deep kernel machines for structured data.** Deep kernel machines combine advantages of kernel methods and deep learning. Both approaches rely on high-dimensional models. Kernels implicitly operate in a space of possibly infinite dimension, whereas deep networks explicitly construct high-dimensional nonlinear data representations. Yet, these approaches are complementary: Kernels can be built with deep learning principles such as hierarchies and convolutions, and approximated by multilayer neural networks. Furthermore, kernels work with structured data and have well understood theoretical principles. Thus, a goal of the Thoth project-team is to design and optimize the training of such deep kernel machines.
- **Large-scale parallel optimization.** Deep kernel machines produce nonlinear representations of input data points. After encoding these data points, a learning task is often formulated as a *large-scale convex optimization problem*; for example, this is the case for linear support vector machines, logistic regression classifiers, or more generally many empirical risk minimization formulations. We intend to pursue recent efforts for making convex optimization techniques that are dedicated to machine learning more scalable. Most existing approaches address scalability issues either in model size (meaning that the function to minimize is defined on a domain of very high dimension), or in the amount of training data (typically, the objective is a large sum of elementary functions). There is

thus a large room for improvements for techniques that jointly take these two criteria into account.

- **Large-scale graphical models.** To represent structured data, we will also investigate graphical models and their optimization. The challenge here is two-fold: designing an adequate cost function and minimizing it. While several cost functions are possible, their utility will be largely determined by the efficiency and the effectiveness of the optimization algorithms for solving them. It is a combinatorial optimization problem involving billions of variables and is NP-hard in general, requiring us to go beyond the classical approximate inference techniques. The main challenges in minimizing cost functions stem from the large number of variables to be inferred, the inherent structure of the graph induced by the interaction terms (e.g., pairwise terms), and the high-arity terms which constrain multiple entities in a graph.

3.4. Datasets and evaluation

Standard benchmarks with associated evaluation measures are becoming increasingly important in computer vision, as they enable an objective comparison of state-of-the-art approaches. Such datasets need to be relevant for real-world application scenarios; challenging for state-of-the-art algorithms; and large enough to produce statistically significant results.

A decade ago, small datasets were used to evaluate relatively simple tasks, such as for example interest point matching and detection. Since then, the size of the datasets and the complexity of the tasks gradually evolved. An example is the Pascal Visual Object Challenge with 20 classes and approximately 10,000 images, which evaluates object classification and detection. Another example is the ImageNet challenge, including thousands of classes and millions of images. In the context of video classification, the TrecVid Multimedia Event Detection challenges, organized by NIST, evaluate activity classification on a dataset of over 200,000 video clips, representing more than 8,000 hours of video, which amounts to 11 months of continuous video.

Almost all of the existing image and video datasets are annotated by hand; it is the case for all of the above cited examples. In some cases, they present limited and unrealistic viewing conditions. For example, many images of the ImageNet dataset depict upright objects with virtually no background clutter, and they may not capture particularly relevant visual concepts: most people would not know the majority of subcategories of snakes cataloged in ImageNet. This holds true for video datasets as well, where in addition a taxonomy of action and event categories is missing.

Our effort on data collection and evaluation will focus on two directions. First, we will design and assemble video datasets, in particular for action and activity recognition. This includes defining relevant taxonomies of actions and activities. Second, we will provide data and define evaluation protocols for weakly supervised learning methods. This does not mean of course that we will forsake human supervision altogether: some amount of ground-truth labeling is necessary for experimental validation and comparison to the state of the art. Particular attention will be payed to the design of efficient annotation tools.

Not only do we plan to collect datasets, but also to provide them to the community, together with accompanying evaluation protocols and software, to enable a comparison of competing approaches for action recognition and large-scale weakly supervised learning. Furthermore, we plan to set up evaluation servers together with leaderboards, to establish an unbiased state of the art on held out test data for which the ground-truth annotations are not distributed. This is crucial to avoid tuning the parameters for a specific dataset and to guarantee a fair evaluation.

- **Action recognition.** We will develop datasets for recognizing human actions and human-object interactions (including multiple persons) with a significant number of actions. Almost all of today's action recognition datasets evaluate classification of short video clips into a number of predefined categories, in many cases a number of different sports, which are relatively easy to identify by their characteristic motion and context. However, in many real-world applications the goal is to identify and localize actions in entire videos, such as movies or surveillance videos of several hours. The actions targeted here are "real-world" and will be defined by compositions of atomic actions into higher-level activities. One essential component is the definition of relevant taxonomies of actions

and activities. We think that such a definition needs to rely on a decomposition of actions into poses, objects and scenes, as determining all possible actions without such a decomposition is not feasible. We plan to provide annotations for spatio-temporal localization of humans as well as relevant objects and scene parts for a large number of actions and videos.

- **Weakly supervised learning.** We will collect weakly labeled images and videos for training. The collection process will be semi-automatic. We will use image or video search engines such as Google Image Search, Flickr or YouTube to find visual data corresponding to the labels. Initial datasets will be obtained by manually correcting whole-image/video labels, i.e., the approach will evaluate how well the object model can be learned if the entire image or video is labeled, but the object model has to be extracted automatically. Subsequent datasets will features noisy and incorrect labels. Testing will be performed on PASCAL VOC'07 and ImageNet, but also on more realistic datasets similar to those used for training, which we develop and manually annotate for evaluation. Our dataset will include both images and videos, the categories represented will include objects, scenes as well as human activities, and the data will be presented in realistic conditions.
- **Joint learning from visual information and text.** Initially, we will use a selection from the large number of movies and TV series for which scripts are available on-line, see for example http://www.dailyscript.com and http://www.weeklyscript.com. These scripts can easily be aligned with the videos by establishing correspondences between script words and (timestamped) spoken ones obtained from the subtitles or audio track. The goal is to jointly learn from visual content and text. To measure the quality of such a joint learning, we will manually annotate some of the videos. Annotations will include the space-time locations of the actions as well as correct parsing of the sentence. While DVDs will, initially, receive most attention, we will also investigate the use of data obtained from web pages, for example images with captions, or images and videos surrounded by text. This data is by nature more noisy than scripts.

TITANE Project-Team

3. Research Program

3.1. Context

Geometric modeling and processing revolve around three main end goals: a computerized shape representation that can be visualized (creating a realistic or artistic depiction), simulated (anticipating the real) or realized (manufacturing a conceptual or engineering design). Aside from the mere editing of geometry, central research themes in geometric modeling involve conversions between physical (real), discrete (digital), and mathematical (abstract) representations. Going from physical to digital is referred to as shape acquisition and reconstruction; going from mathematical to discrete is referred to as shape approximation and mesh generation; going from discrete to physical is referred to as shape rationalization.

Geometric modeling has become an indispensable component for computational and reverse engineering. Simulations are now routinely performed on complex shapes issued not only from computer-aided design but also from an increasing amount of available measurements. The scale of acquired data is quickly growing: we no longer deal exclusively with individual shapes, but with entire *scenes*, possibly at the scale of entire cities, with many objects defined as structured shapes. We are witnessing a rapid evolution of the acquisition paradigms with an increasing variety of sensors and the development of community data, as well as disseminated data.

In recent years, the evolution of acquisition technologies and methods has translated in an increasing overlap of algorithms and data in the computer vision, image processing, and computer graphics communities. Beyond the rapid increase of resolution through technological advances of sensors and methods for mosaicing images, the line between laser scan data and photos is getting thinner. Combining, e.g., laser scanners with panoramic cameras leads to massive 3D point sets with color attributes. In addition, it is now possible to generate dense point sets not just from laser scanners but also from photogrammetry techniques when using a well-designed acquisition protocol. Depth cameras are getting increasingly common, and beyond retrieving depth information we can enrich the main acquisition systems with additional hardware to measure geometric information about the sensor and improve data registration: e.g., accelerometers or GPS for geographic location, and compasses or gyrometers for orientation. Finally, complex scenes can be observed at different scales ranging from satellite to pedestrian through aerial levels.

These evolutions allow practitioners to measure urban scenes at resolutions that were until now possible only at the scale of individual shapes. The related scientific challenge is however more than just dealing with massive data sets coming from increase of resolution, as complex scenes are composed of multiple objects with structural relationships. The latter relate i) to the way the individual shapes are grouped to form objects, object classes or hierarchies, ii) to geometry when dealing with similarity, regularity, parallelism or symmetry, and iii) to domain-specific semantic considerations. Beyond reconstruction and approximation, consolidation and synthesis of complex scenes require rich structural relationships.

The problems arising from these evolutions suggest that the strengths of geometry and images may be combined in the form of new methodological solutions such as photo-consistent reconstruction. In addition, the process of measuring the geometry of sensors (through gyrometers and accelerometers) often requires both geometry process and image analysis for improved accuracy and robustness. Modeling urban scenes from measurements illustrates this growing synergy, and it has become a central concern for a variety of applications ranging from urban planning to simulation through rendering and special effects.

3.2. Analysis

Complex scenes are usually composed of a large number of objects which may significantly differ in terms of complexity, diversity, and density. These objects must be identified and their structural relationships must be recovered in order to model the scenes with improved robustness, low complexity, variable levels of details and ultimately, semantization (automated process of increasing degree of semantic content).

Object classification is an ill-posed task in which the objects composing a scene are detected and recognized with respect to predefined classes, the objective going beyond scene segmentation. The high variability in each class may explain the success of the stochastic approach which is able to model widely variable classes. As it requires a priori knowledge this process is often domain-specific such as for urban scenes where we wish to distinguish between instances as ground, vegetation and buildings. Additional challenges arise when each class must be refined, such as roof super-structures for urban reconstruction.

Structure extraction consists in recovering structural relationships between objects or parts of object. The structure may be related to adjacencies between objects, hierarchical decomposition, singularities or canonical geometric relationships. It is crucial for effective geometric modeling through levels of details or hierarchical multiresolution modeling. Ideally we wish to learn the structural rules that govern the physical scene manufacturing. Understanding the main canonical geometric relationships between object parts involves detecting regular structures and equivalences under certain transformations such as parallelism, orthogonality and symmetry. Identifying structural and geometric repetitions or symmetries is relevant for dealing with missing data during data consolidation.

Data consolidation is a problem of growing interest for practitioners, with the increase of heterogeneous and defect-laden data. To be exploitable, such defect-laden data must be consolidated by improving the data sampling quality and by reinforcing the geometrical and structural relations sub-tending the observed scenes. Enforcing canonical geometric relationships such as local coplanarity or orthogonality is relevant for registration of heterogeneous or redundant data, as well as for improving the robustness of the reconstruction process.

3.3. Approximation

Our objective is to explore the approximation of complex shapes and scenes with surface and volume meshes, as well as on surface and domain tiling. A general way to state the shape approximation problem is to say that we search for the shape discretization (possibly with several levels of detail) that realizes the best complexity / distortion trade-off. Such a problem statement requires defining a discretization model, an error metric to measure distortion as well as a way to measure complexity. The latter is most commonly expressed in number of polygon primitives, but other measures closer to information theory lead to measurements such as number of bits or minimum description length.

For surface meshes we intend to conceive methods which provide control and guarantees both over the global approximation error and over the validity of the embedding. In addition, we seek for resilience to heterogeneous data, and robustness to noise and outliers. This would allow repairing and simplifying triangle soups with cracks, self-intersections and gaps. Another exploratory objective is to deal generically with different error metrics such as the symmetric Hausdorff distance, or a Sobolev norm which mixes errors in geometry and normals.

For surface and domain tiling the term meshing is substituted for tiling to stress the fact that tiles may be not just simple elements, but can model complex smooth shapes such as bilinear quadrangles. Quadrangle surface tiling is central for the so-called *resurfacing* problem in reverse engineering: the goal is to tile an input raw surface geometry such that the union of the tiles approximates the input well and such that each tile matches certain properties related to its shape or its size. In addition, we may require parameterization domains with a simple structure. Our goal is to devise surface tiling algorithms that are both reliable and resilient to defect-laden inputs, effective from the shape approximation point of view, and with flexible control upon the structure of the tiling.

3.4. Reconstruction

Assuming a geometric dataset made out of points or slices, the process of shape reconstruction amounts to recovering a surface or a solid that matches these samples. This problem is inherently ill-posed as infinitely-many shapes may fit the data. One must thus regularize the problem and add priors such as simplicity or smoothness of the inferred shape.

The concept of geometric simplicity has led to a number of interpolating techniques commonly based upon the Delaunay triangulation. The concept of smoothness has led to a number of approximating techniques that commonly compute an implicit function such that one of its isosurfaces approximates the inferred surface. Reconstruction algorithms can also use an explicit set of prior shapes for inference by assuming that the observed data can be described by these predefined prior shapes. One key lesson learned in the shape problem is that there is probably not a single solution which can solve all cases, each of them coming with its own distinctive features. In addition, some data sets such as point sets acquired on urban scenes are very domain-specific and require a dedicated line of research.

In recent years the *smooth, closed case* (i.e., shapes without sharp features nor boundaries) has received considerable attention. However, the state-of-the-art methods have several shortcomings: in addition to being in general not robust to outliers and not sufficiently robust to noise, they often require additional attributes as input, such as lines of sight or oriented normals. We wish to devise shape reconstruction methods which are both geometrically and topologically accurate without requiring additional attributes, while exhibiting resilience to defect-laden inputs. Resilience formally translates into stability with respect to noise and outliers. Correctness of the reconstruction translates into convergence in geometry and (stable parts of) topology of the reconstruction with respect to the inferred shape known through measurements.

Moving from the smooth, closed case to the *piecewise smooth case* (possibly with boundaries) is considerably harder as the ill-posedness of the problem applies to each sub-feature of the inferred shape. Further, very few approaches tackle the combined issue of robustness (to sampling defects, noise and outliers) and feature reconstruction.

TYREX Project-Team

3. Research Program

3.1. Foundations for Data Manipulation Analysis: Logics and Type Systems

We develop methods for the static analysis of queries based on logical decision procedures. Static analysis can be used to optimize runtime performance by compile-time automated modification of the code. For example, queries can be substituted by more efficient — yet equivalent — variants. The query containment problem has been a central point of research for major query languages due to its vital role in query optimization. Query containment is defined as determining if the result of one query is included in the result of another one for any dataset. We explore techniques for deciding query containment for expressive languages for querying richly structured data such as knowledge graphs. One major scientific difficulty here consists in dealing with problems close to the frontier of decidability, and therefore in finding useful trade-offs between programming expressivity, complexity, succinctness, algorithmic techniques and effective implementations. We also investigate type systems and type-checking methods for the analysis of the manipulations of structured data.

3.2. Algebraic Foundations for Query Optimization and Code Synthesis

We consider intermediate languages based on algebraic foundations for the representation, characterization, transformations and compilation of queries. We investigate extensions of the relational algebra for optimizing expressive queries, and in particular recursive queries. We explore monads and in particular monad comprehensions and monoid calculus for the generation of efficient and scalable code on big data frameworks. When transforming and optimizing algebraic terms, we rely on cost-based searches of equivalent terms. We thus develop cost models whose purpose is to estimate the time, space and network costs of query evaluation. One difficulty is to estimate these costs in architectures where data and computations are distributed, and where the modeling of data transfers is essential.

VALDA Project-Team

3. Research Program

3.1. Scientific Foundations

We now detail some of the scientific foundations of our research on complex data management. This is the occasion to review connections between data management, especially on complex data as is the focus of Valda, with related research areas.

3.1.1. Complexity & Logic

Data management has been connected to logic since the advent of the relational model as main representation system for real-world data, and of first-order logic as the logical core of database querying languages [37]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [63], recursive queries (Datalog), or querying of XML databases [46]. Logical formalisms facilitate reasoning about the expressiveness of a query language or about its complexity.

The main problem of interest in data management is that of query evaluation, i.e., computing the results of a query over a database. The complexity of this problem has far-reaching consequences. For example, it is because first-order logic is in the AC_0 complexity class that evaluation of SQL queries can be parallelized efficiently. It is usual [74] in data management to distinguish *data complexity*, where the query is considered to be fixed, from *combined complexity*, where both the query and the data are considered to be part of the input. Thus, though conjunctive queries, corresponding to a simple SELECT-FROM-WHERE fragment of SQL, have PTIME data complexity, they are NP-hard in combined complexity. Making this distinction is important, because data is often far larger (up to the order of terabytes) than queries (rarely more than a few hundred bytes). Beyond simple query evaluation, a central question in data management remains that of complexity; tools from algorithm analysis, and complexity theory can be used to pinpoint the tractability frontier of data management tasks.

3.1.2. Automata Theory

Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [36], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [42], queries are naturally given as automata; typical query and schema languages for XML databases such as XPath and XML Schema can be compiled to tree automata [67], or for more complex languages to data tree automata[4]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [50] that show that very expressive queries (from the language of monadic second-order language) can be evaluated as tree automata over *tree decompositions* of the original databases, yielding linear-time algorithms (in data complexity) for a wide variety of applications.

3.1.3. Verification

Complex data management also has connections to verification and static analysis. Besides query evaluation, a central problem in data management is that of deciding whether two queries are *equivalent* [37]. This is critical for query optimization, in order to determine if the rewriting of a query, maybe cheaper to evaluate, will return the same result as the original query. Equivalence can easily be seen to be an instance of the problem of (non-)satisfiability: $q \equiv q'$ if and only if $(q \wedge \neg q') \vee (\neg q \wedge q')$ is not satisfiable. In other words, some aspects of query optimization are static analysis issues. Verification is also a critical part of any database application where it is important to ensure that some property will never (or always) arise [48].

3.1.4. Workflows

The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of data management applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. Different approaches to workflow specifications exist: automata-based, logic-based, or predicate-based control of function calls [34].

3.1.5. Probability & Provenance

To deal with the uncertainty attached to data, proper models need to be used (such as attaching *provenance* information to data items and viewing the whole database as being *probabilistic*) and practical methods and systems need to be developed to both reliably estimate the uncertainty in data items and properly manage provenance and uncertainty information throughout a long, complex system.

The simplest model of data uncertainty is the NULLs of SQL databases, also called Codd tables [37]. This representation system is too basic for any complex task, and has the major inconvenient of not being closed under even simple queries or updates. A solution to this has been proposed in the form of *conditional tables* [61] where every tuple is annotated with a Boolean formula over independent Boolean random events. This model has been recognized as foundational and extended in two different directions: to more expressive models of *provenance* than what Boolean functions capture, through a semiring formalism [57], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [58]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [49], [43]. Research in the past ten years has focused on a better understanding of the tractability of query answering with provenance and probabilistic annotations, in a variety of specializations of this framework [72] [62], [40].

3.1.6. Machine Learning

Statistical machine learning, and its applications to data mining and data analytics, is a major foundation of data management research. A large variety of research areas in complex data management, such as wrapper induction [68], crowdsourcing [41], focused crawling [56], or automatic database tuning [44] critically rely on machine learning techniques, such as classification [60], probabilistic models [55], or reinforcement learning [73].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [65] system result in probabilistic annotations that need to be properly modeled, stored, and queried.

Finally, complex data management also brings new twists to some classical machine learning problems. Consider for instance the area of *active learning* [70], a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. In most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [69] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [53]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

3.2. Research Directions

At the beginning of the Valda team, the project was to focus on the following directions:

- foundational aspects of data management, in particular related to query enumeration and reasoning on data, especially regarding security issues;
- implementation of provenance and uncertainty management, real-world applications, other aspects of uncertainty and incompleteness, in particular dynamic;
- development of personal information management systems, integration of machine learning techniques.

We believe the first two directions have been followed in a satisfactory manner. The focus on personal information management has not been kept for various organizational reasons, however, but the third axis of the project is reoriented to more general aspects of Web data management.

New permanent arrivals in the group since its creation have impacted its research directions in the following manner:

- Camille BOURGAUX and Michaël THOMAZO are both specialists of knowledge representation and formal aspects of knowledge bases, which is an expertise that did not exist in the group. They are also both interested in, and have started working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This will lead to more work on knowledge representation and symbolic AI aspects, while keeping the focus of Valda on foundations of data management and uncertainty.
- Olivier CAPPÉ is a specialist in statistics and machine learning, in particular multi-armed bandits and reinforcement learning. He is also interested in applications of these learning techniques to data management problems. His arrival in the group therefore complements the expertise of other researchers, and will lead to more work on machine learning issues.
- Leonid LIBKIN is a specialist of database theory, of incomplete data management, and has a line of current research on graph data management. His profile fits very well with the original orientation of the Valda project.

We intend to keep producing leading research on the foundations of data management. Generally speaking, the goal is to investigate the borders of feasibility of various tasks. For instance, what are the assumptions on data that allow for computable problems? When is it not possible at all? When can we hope for efficient query answering, when is it hopeless? This is a problem of theoretical nature which is necessary for understanding the limit of the methods and driving research towards the scenarios where positive results may be obtainable. Only when we have understood the limitation of different methods and have many examples where this is possible, we can hope to design a solid foundation that allowing for a good trade-off between what can be done (needs from the users) and what can be achieved (limitation from the system).

Similarly, we will continue our work, both foundational and practical, on various aspects of provenance and uncertainty management. One overall long-term goal is to reach a full understanding of the interactions between query evaluation or other broader data management tasks and uncertain and annotated data models. We would in particular want to go towards a full classification of tractable (typically polynomial-time) and intractable (typically NP-hard for decision problems, or #P-hard for probability evaluation) tasks, extending and connecting the query-based dichotomy [51] on probabilistic query evaluation with the instance-based one of [39], [40]. Another long-term goal is to consider more dynamic scenarios than what has been considered so far in the uncertain data management literature: when following a workflow, or when interacting with intensional data sources, how to properly represent and update uncertainty annotations that are associated with data. This is critical for many complex data management scenarios where one has to maintain a probabilistic current knowledge of the world, while obtaining new knowledge by posing queries and accessing data sources. Such intensional tasks requires minimizing jointly data uncertainty and cost to data access.

As application area, in addition to the historical focus on personal information management which is now less stressed, we target Web data (Web pages, the semantic Web, social networks, the deep Web, crowdsourcing platforms, etc.).

We aim at keeping a delicate balance between theoretical, foundational research, and systems research, including development and implementation. This is a difficult balance to find, especially since most Valda researchers have a tendency to favor theoretical work, but we believe it is also one of the strengths of the team.

WILLOW Team

3. Research Program

3.1. 3D object and scene modeling, analysis, and retrieval

This part of our research focuses on geometric models of specific 3D objects at the local (differential) and global levels, physical and statistical models of materials and illumination patterns, and modeling and retrieval of objects and scenes in large image collections. Our past work in these areas includes research aimed at recognizing rigid 3D objects in cluttered photographs taken from arbitrary viewpoints (Rothganger *et al.*, 2006), segmenting video sequences into parts corresponding to rigid scene components before recognizing these in new video clips (Rothganger *et al.*, 2007), retrieval of particular objects and buildings from images and videos (Sivic and Zisserman, 2003) and (Philbin *et al.*, 2007), and a theoretical study of a general formalism for modeling central and non-central cameras using the formalism and terminology of classical projective geometry (Ponce, 2009 and Batog *et al.*, 2010).

We have also developed multi-view stereopsis algorithms that have proven remarkably effective at recovering intricate details and thin features of compact objects and capturing the overall structure of large-scale, cluttered scenes. We have obtained a US patent 8,331,615 [0] for the corresponding software (PMVS, https://github.com/pmoulon/CMVS-PMVS) which is available under a GPL license and used for film production by ILM and Weta as well as by Google in Google Maps. It is also the basic technology used by Iconem, a start-up founded by Y. Ubelmann, a Willow collaborator. We have also applied our multi-view-stereo approach to model archaeological sites together with developing representations and efficient retrieval techniques to enable matching historical paintings to 3D models of archaeological sites (Russel *et al.*, 2011).

Our current efforts in this area are outlined in detail in Section. 7.1 .

3.2. Category-level object and scene recognition

The objective in this core part of our research is to learn and recognize quickly and accurately thousands of visual categories, including materials, objects, scenes, and broad classes of temporal events, such as patterns of human activities in picnics, conversations, etc. The current paradigm in the vision community is to model/learn one object category (read 2D aspect) at a time. If we are to achieve our goal, we have to break away from this paradigm, and develop models that account for the tremendous variability in object and scene appearance due to texture, material, viewpoint, and illumination changes within each object category, as well as the complex and evolving relationships between scene elements during the course of normal human activities.

Our current work in this area is outlined in detail in Section 7.2 .

3.3. Image restoration, manipulation and enhancement

The goal of this part of our research is to develop models, and methods for image/video restoration, manipulation and enhancement. The ability to “intelligently" manipulate the content of images and video is just as essential as high-level content interpretation in many applications: This ranges from restoring old films or removing unwanted wires and rigs from new ones in post production, to cleaning up a shot of your daughter at her birthday party, which is lovely but noisy and blurry because the lights were out when she blew the candles, or editing out a tourist from your Roman holiday video. Going beyond the modest abilities of current “digital zoom" (bicubic interpolation in general) so you can close in on that birthday cake, “deblock" a football game on TV, or turn your favorite DVD into a blue-ray, is just as important.

[0] The patent: "Match, Expand, and Filter Technique for Multi-View Stereopsis" was issued December 11, 2012 and assigned patent number 8,331,615.

In this context, we believe there is a new convergence between computer vision, machine learning, and signal processing. For example: The idea of exploiting self-similarities in image analysis, originally introduced in computer vision for texture synthesis applications (Efros and Leung, 1999), is the basis for non-local means (Buades *et al.*, 2005), one of today's most successful approaches to image restoration. In turn, by combining a powerful sparse coding approach to non-local means (Dabov *et al.*, 2007) with modern machine learning techniques for dictionary learning (Mairal *et al.*, 2010), we have obtained denoising and demosaicking results that are the state of the art on standard benchmarks (Mairal *et al.*, 2009).

Our current work is outlined in detail in Section 7.3 .

3.4. Human activity capture and classification

From a scientific point of view, visual action understanding is a computer vision problem that until recently has received little attention outside of extremely specific contexts such as surveillance or sports. Many of the current approaches to the visual interpretation of human activities are designed for a limited range of operating conditions, such as static cameras, fixed scenes, or restricted actions. The objective of this part of our project is to attack the much more challenging problem of understanding actions and interactions in unconstrained video depicting everyday human activities such as in sitcoms, feature films, or news segments. The recent emergence of automated annotation tools for this type of video data (Everingham, Sivic, Zisserman, 2006; Laptev, Marszałek, Schmid, Rozenfeld, 2008; Duchenne, Laptev, Sivic, Bach, Ponce, 2009) means that massive amounts of labelled data for training and recognizing action models will at long last be available.

Our research agenda in this scientific domain is described below and our recent results are outlined in detail in Section 7.4 .

- **Weakly-supervised learning and annotation of human actions in video.** We aim to leverage the huge amount of video data using readily-available annotations in the form of video scripts. Scripts, however, often provide only imprecise and incomplete information about the video. We address this problem with weakly-supervised learning techniques both at the text and image levels.
- **Descriptors for video representation.** Video representation has a crucial role for recognizing human actions and other components of a visual scene. Our work in this domain aims to develop generic methods for representing video data based on realistic assumptions. In particular, we develop deep learning methods and design new trainable representations for various tasks such as human action recognition, person detection, segmentation and tracking.

3.5. Learning embodied representations

Computer vision has come a long way toward understanding images and videos in terms of scene geometry, object labels, locations and poses of people or classes of human actions. This "understanding", however, remains largely disconnected from reasoning about the physical world. For example, what will happen if removing a tablecloth from a setted table? What actions will be needed to resume an interrupted meal? We believe that a true *embodied* understanding of dynamic scenes from visual observations is the next major research challenge. We plan to address this challenge by developing new models and algorithms with an emphasis on the synergy between vision, learning, robotics and natural language understanding. If successful, this research direction will bring significant advances in high-impact applications such as autonomous driving, home robotics and personal visual assistance.

Learning embodied representations is planned to be a major research axis for the successor of the Willow team. Meanwhile we have already started work in this direction and report our first results in Section 7.5 .

WIMMICS Project-Team

3. Research Program

3.1. Users Modeling and Designing Interaction on the Web

Wimmics focuses on interactions of ordinary users with ontology-based knowledge systems, with a preference for semantic Web formalisms and Web 2.0 applications. We specialize interaction design and evaluation methods to Web application tasks such as searching, browsing, contributing or protecting data. The team is especially interested in using semantics in assisting the interactions. We propose knowledge graph representations and algorithms to support interaction adaptation, for instance for context-awareness or intelligent interactions with machine. We propose and evaluate Web-based visualization techniques for linked data, querying, reasoning, explaining and justifying. Wimmics also integrates natural language processing approaches to support natural language based interactions. We rely on cognitive studies to build models of the system, the user and the interactions between users through the system, in order to support and improve these interactions. We extend the user modeling technique known as *Personas* where user models are represented as specific, individual humans. *Personas* are derived from significant behavior patterns (i.e., sets of behavioral variables) elicited from interviews with and observations of users (and sometimes customers) of the future product. Our user models specialize *Personas* approaches to include aspects appropriate to Web applications. Wimmics also extends user models to capture very different aspects (e.g. emotional states).

3.2. Communities and Social Interactions Analysis

The domain of social network analysis is a whole research domain in itself and Wimmics targets what can be done with typed graphs, knowledge representations and social models. We also focus on the specificity of social Web and semantic Web applications and in bridging and combining the different social Web data structures and semantic Web formalisms. Beyond the individual user models, we rely on social studies to build models of the communities, their vocabularies, activities and protocols in order to identify where and when formal semantics is useful. We propose models of collectives of users and of their collaborative functioning extending the collaboration personas and methods to assess the quality of coordination interactions and the quality of coordination artifacts. We extend and compare community detection algorithms to identify and label communities of interest with the topics they share. We propose mixed representations containing social semantic representations (e.g. folksonomies) and formal semantic representations (e.g. ontologies) and propose operations that allow us to couple them and exchange knowledge between them. Moving to social interaction we develop models and algorithms to mine and integrate different yet linked aspects of social media contributions (opinions, arguments and emotions) relying in particular on natural language processing and argumentation theory. To complement the study of communities we rely on multi-agent systems to simulate and study social behaviors. Finally we also rely on Web 2.0 principles to provide and evaluate social Web applications.

3.3. Vocabularies, Semantic Web and Linked Data Based Knowledge Representation and Artificial Intelligence Formalisms on the Web

For all the models we identified in the previous sections, we rely on and evaluate knowledge representation methodologies and theories, in particular ontology-based modeling. We also propose models and formalisms to capture and merge representations of different levels of semantics (e.g. formal ontologies and social folksonomies). The important point is to allow us to capture those structures precisely and flexibly and yet create as many links as possible between these different objects. We propose vocabularies and semantic Web formalizations for all the aspects that we model and we consider and study extensions of these formalisms when needed. The results have all in common to pursue the representation and publication of our models

as linked data. We also contribute to the transformation and linking of existing resources (informal models, databases, texts, etc.) to be published on the Semantic Web and as Linked Data. Examples of aspects we formalize include: user profiles, social relations, linguistic knowledge, business processes, derivation rules, temporal descriptions, explanations, presentation conditions, access rights, uncertainty, emotional states, licenses, learning resources, etc. At a more conceptual level we also work on modeling the Web architecture with philosophical tools so as to give a realistic account of identity and reference and to better understand the whole context of our research and its conceptual cornerstones.

3.4. Artificial Intelligence Processing: Learning, Analyzing and Reasoning on Heterogeneous Semantic Graphs

One of the characteristics of Wimmics is to rely on graph formalisms unified in an abstract graph model and operators unified in an abstract graph machine to formalize and process semantic Web data, Web resources, services metadata and social Web data. In particular Corese, the core software of Wimmics, maintains and implements that abstraction. We propose algorithms to process the mixed representations of the previous section. In particular we are interested in allowing cross-enrichment between them and in exploiting the life cycle and specificity of each one to foster the life-cycles of the others. Our results all have in common to pursue analyzing and reasoning on heterogeneous semantic graphs issued from social and semantic Web applications. Many approaches emphasize the logical aspect of the problem especially because logics are close to computer languages. We defend that the graph nature of Linked Data on the Web and the large variety of types of links that compose them call for typed graphs models. We believe the relational dimension is of paramount importance in these representations and we propose to consider all these representations as fragments of a typed graph formalism directly built above the Semantic Web formalisms. Our choice of a graph based programming approach for the semantic and social Web and of a focus on one graph based formalism is also an efficient way to support interoperability, genericity, uniformity and reuse.

ZENITH Project-Team

3. Research Program

3.1. Distributed Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMS, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (search engines, application servers, document systems, etc.).

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL). Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved to be effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases.

Parallel database systems extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

In contrast, peer-to-peer (P2P) systems adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. P2P systems typically have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g. file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. A P2P solution is well-suited to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources. But for very-large scale scientific data analysis, we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the best of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.2. Big Data

Big data (like its relative, data science) has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte of Hard Disk Drive for: 1M$ in 1982, 1K$ in 1995, 0.02$ in 2015), making it affordable to keep more data around. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's such as: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Many different big data management solutions have been designed, primarily for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility, hence the new term Data-Intensive Scalable Computing (DISC). Examples of DISC systems include data processing frameworks (e.g. Hadoop MapReduce, Apache Spark, Pregel), file systems (e.g. Google GFS, HDFS), NoSQL systems (Google BigTable, Hbase, MongoDB), NewSQL systems (Google F1, CockroachDB, LeanXcale). In Zenith, we exploit or extend DISC technologies to fit our needs for scientific workflow management and scalable data analysis.

3.3. Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are integrated and materialized in a data warehouse. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

In both cases, to understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the autonomy of the underlying data sources, which leads to a large variety of models and formats. Thus, it is necessary to identify semantic correspondences between the metadata of the related data sources. This requires the matching of the heterogeneous metadata, by discovering semantic correspondences between ontologies, and the annotation of data sources using ontologies. In Zenith, we rely on semantic web techniques (e.g. RDF and SparkQL) to perform these tasks and deal with high numbers of data sources.

Scientific workflow management systems (SWfMS) are also useful for data integration. They allow scientists to describe and execute complex scientific activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data. This requires using distributed and parallel execution environments. However, existing workflow management systems have limited support for data parallelism. In Zenith, we use an algebraic approach to describe data-intensive workflows and exploit parallelism.

3.4. Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management, and is applied to categorical and continuous data. In the Zenith team, we are interested in both of these data types. Categorical data designates a set of data that can be described as "check boxes". It can be names, products, items, towns, etc. A common illustration is the market basket data, where each item bought by a client is recorded and the set of items is the basket. The typical data mining problems with this kind of data are:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset discovery but considering the order between. In the smart building example, a frequent sequence could say that "in 40% of rooms, lights are on at time i, the room is empty at time $i + j$ and the door is closed at time $i + j + k$".
- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

Continuous data are numeric records that can have an infinite number of values between any two values. A temperature value or a timestamp are examples of such data. They are involved in a widely used type of data known as time series: a series of values, ordered by time, and giving a measure, e.g. coming from a sensor. There is a large number of problems that can apply to this kind of data, including:

- **Indexing and retrieval.** The goal, here, is usually to find, given a query q and a time series dataset D, the records of D that are most similar to q. This may involve any transformation of D by means of an index or an alternative representation for faster execution.
- **Pattern and outlier detection.** The discovery of recurrent patterns or atypical sub-windows in a time series has applications in finance, industrial manufacture or seismology, to name a few. It calls for techniques that avoid pairwise comparisons of all the sub-windows, which would lead to prohibitive response times.
- **Clustering**. The goal is the same as categorical data clustering: group similar time series and separate dissimilar ones.

One main problem in data analytics is to deal with data streams. Existing methods have been designed for very large data sets where complex algorithms from artificial intelligence were not efficient because of data size. However, we now must deal with data streams, sequences of data events arriving at high rate, where traditional data analytics techniques cannot complete in real-time, given the infinite data size. In order to extract knowledge from data streams, the data mining community has investigated approximation methods that could yield good result quality.

3.5. High dimensional data processing and search

High dimensionality is inherent in applications involving images, audio and text as well as in many scientific applications involving raster data or high-throughput data. Because of the *dimensionality curse*, technologies for processing and analyzing such data cannot rely on traditional relational DBMS or data mining methods. It rather requires to employ machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods that permit data processing and search at scale, in particular in the presence of strong uncertainty and/or ambiguity. Actually, while small datasets are often characterized by a careful collection process, massive amounts of data often come with outliers and spurrious items, because it appears impossible to guarantee faultless collection at massive

bandwidth. Another source of noise is often the sensor itself, that may be of low quality but of high sampling rate, or even the actual content, e.g. in cultural heritage applications when historical content appears seriously damaged by time. To attack these difficult problems, we focus on the following research topics:

- **Uncertainty estimation**. Items in massive datasets may either be uncertain, e.g. for automatically annotated data as in image analysis, or be more or less severely corrupted by noise, e.g. in noisy audio recordings or in the presence of faulty sensors. In both cases, the concept of *uncertainty* is central for the end-user to exploit the content. In this context, we use probability theory to quantify uncertainty and propose machine learning algorithms that may operate robustly, or at least assess the quality of their output. This vast topic of research is guided by large-scale applications (both data search and data denoising), and our research is oriented towards computationally effective methods.
- **Deep neural networks**. A major breakthrough in machine learning performance has been the advent of deep neural networks, which are characterized by high numbers (millions) of parameters and scalable learning procedures. We are striving towards original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and processing. The specific challenges we investigate are: very high dimensionality for static data and very long-term dependency for temporal data, both in the case of possibly strong uncertainty or ambiguity (e.g. hundreds of thousands of classes).
- **Community service**. Research in machine learning is guided by applications. In Zenith, two main communities are targeted: botany, and digital humanities. In both cases, our observation is that significant breakthroughs may be achieved by connecting these communities to machine learning researchers. This may be achieved through wording application-specific problems in classical machine learning parlance. Thus, the team is actively involved in the organization of international evaluation campaigns that allow machine learning researchers to propose new methods while solving important application problems.