RESEARCH CENTER
Paris

FIELD

Activity Report 2018

Section Scientific Foundations

Edition: 2019-03-07

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* [27]. 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* [27], 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 [26], 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.

AOSTE2 Team

3. Research Program

3.1. The Algorithm-Architecture Adequation methodology and Real-Time Scheduling

Participants: Liliana Cucu, Dumitru Potop Butucaru, Yves Sorel.

The Algorithm-Architecture Adequation (AAA) methodology relies on distributed real-time schedulability and optimization theories to map efficiently an algorithm model to an architecture model.

The algorithm model which describes the functional specifications of the applications, is an extension of the well known data-flow model from Dennis [14]. It is a directed acyclic hyper-graph (DAG) that we call "conditioned factorized data dependence graph", whose vertices are functions and hyper-edges are directed "data or control dependences" between functions. The data dependences define a partial order on the functions execution. The basic data-flow model was extended in three directions: first infinite (resp. finite) repetition of a sub-graph pattern in order to specify the reactive aspect of real-time systems (resp. in order to specify the finite repetition of a sub-graph consuming different data similar to a loop in imperative languages), second "state" when data dependences are necessary between different infinite repetitions of the sub-graph pattern introducing cycles which must be avoided by introducing specific vertices called "delays" (similar to z -n in automatic control), third "conditioning" of a function by a control dependence similar to conditional control structure in imperative languages, allowing the execution of alternative subgraphs. Delays combined with conditioning allow the programmer to specify automata used for describing "mode changes".

The architecture model which describes the non functional specifications is, in the simplest case, a directed graph whose vertices are of two types: "processor" (one sequencer of functions, several sequencers of communications and distributed or shared memories) and "medium" (multiplexers and demultiplexers), and whose edges are directed connections. With such model it is possible to describe classic heterogeneous distributed, parallel and multiprocessor platforms as well as the most recent multi/manycore platforms. The worst case times mentioned previously are estimated according to this model.

The implementation model is a graph obtained by applying an external composition law such that an architecture graph operates on an algorithm graph to give an algorithm graph while taking advantage of timing characteristics, basically periods, deadlines and WCETs. This resulting algorithm graph is built by performing spatial and timing allocations (distribution and scheduling) of algorithm graph functions on architecture graph resources, and of dependences between functions on communication media. In that context "Adequation" means to search, in the solution space of implementation graphs, one implementation graph which verifies real-time constraints and, in addition, minimizes some criteria. These criteria consists in the total execution time of the algorithm executed on the architecture, the number of computing or communication resources, etc. Below, we describe distributed real-time schedulability analyses and optimization techniques suited for that purposes.

We address two main issues: uniprocessor and multiprocessor real-time scheduling for which some real-time constraints are of high criticality, i.e. they must be satisfied otherwise dramatic consequences occur.

In the case of uniprocessor real-time scheduling, besides the usual deadline constraint, often equal to the period of each task, i.e. a function with timing characteristics, we take into consideration dependences beetween tasks, and possibly several latencies. The latter are "end-to-end" constraints that may have complex relationships. Dealing with multiple real-time constraints raises the complexity of the scheduling problems. Moreover, costs of the Real-Time Operating System (RTOS) and of preemptions lead to, at least, a waste of resources due to their approximation in the WCET (Worst Execution Time) of each task, as proposed by Liu and Layland in their seminal article [21]. This is the reason why we first studied non-preemptive real-time scheduling with dependences, periodicities, and latencies constraints. Although a bad approximation of costs of the RTOS and

of preemptions, may have dramatic consequences on real-time scheduling, there are only few researches on this topic. Thus, we investigated preemptive real-time scheduling while taking into account its cost which is very difficult to determine because it varies according to the instance (job) of each task. This latter is integrated in the schedulability conditions, and in the corresponding scheduling algorithms we propose. More generally, we integrate in schedulability analyses costs of the RTOS and of preemptions.

In the case of multiprocessor real-time scheduling, we chose to study first the "partitioned approach", rather than the "global approach", since the latter uses task migrations whose cost is prohibitive for current commercial processors, even for the more recent many/multicore. The partitioned approach enables us to reuse the results obtained in the uniprocessor case in order to derive solutions for the multiprocessor case. We consider also the semi-partitioned approach which allows only some migrations in order to minimize their costs. In addition, to satisfy the multiple real-time constraints mentioned in the uniprocessor case, we have to minimize the total execution time (makespan) since we deal with automatic control applications involving feedback loops. The complexity of such minimization problem increases because the cost of interprocessor communications (through buses in a multi-processor or routers in a manycore) must be taken into account. Furthermore, the domain of embedded systems leads to solving minimization resources problems. Since both optimization problems are NP-hard we develop exact algorithms (ILP, B & B, B & C) which are optimal for simple problems, and heuristics which are sub-optimal for realistic problems corresponding to industrial needs. Long time ago we proposed a very fast "greedy" heuristics whose results were regularly improved, and extended with local neighborhood heuristics, or used as initial solutions for metaheuristics.

Besides the spatial dimension (distributed) of the real-time scheduling problem, other important dimensions are the type of communication mechanisms (shared memory vs. message passing), or the source of control and synchronization (event-driven vs. time-triggered). We explore real-time scheduling on architectures corresponding to all combinations of the above dimensions. This is of particular impact in application domains such as railways and avionics.

3.2. Probabilistic Worst Case Reasoning for Real-Time Systems

Participants: Liliana Cucu, Robert Davis, Yves Sorel.

The arrival of modern hardware responding to the increasing demand for new functionalities exacerbates the limitations of the current worst-case real-time reasoning, mainly to the rarity of worst-case scenarios. Several solutions exist to overcome this important pessimism and our solution takes into account the extremely low probability of appearance of a worst-case scenario within one hour of functioning (10^{-45}), compared to the certification requirements for instance (10^{-9} for the highest level of certification in avionics). Thus we model and analyze real-time systems with time parameters described by using probabilistic models. Our results for such models address both schedulability analyses as well as timing analyses. Both such analyses are impacted by existing misunderstanding. The independence between tasks is a property of real-time systems that is often used for its basic results. Any complex model takes into account different dependences caused by sharing resources other than the processor. On another hand, the probabilistic operations require, generally, the (probabilistic) independence between the random variables describing some parameters of a probabilistic real-time system. The main (original) criticism to probabilistic is based on this hypothesis of independence judged too restrictive to model real-time systems. In reality the two notions of independence are different. Providing arguments to underline this confusion is at the center of our dissemination effort in the last years.

We provide below the bases driving our current research as follows:

- *Optimality of scheduling algorithms* stays an important aspect of the probabilistic real-time systems, especially that the introduction of probabilistic time parameters has a direct impact on the optimality of the existing scheduling algorithms. For instance Rate Monotonic scheduling policy is no longer optimal in the case of one processor when a preemptive fixed-priority solution exists. We expect other classes of algorithms to lose their optimality and we concentrate our efforts to propose new scheduling solutions in this context [22].

- *Increased complexity of schedulability analysis* due to the introduction of probabilistic parameters requires appropriate complexity reasoning, especially with the emergence of probabilistic schedulability analyses for mixed-criticality real-time systems [23]. Moreover the real-time applications are rarely independent and precedence constraint using graph-based models are appropriate in this context. Precedence constraints do decrease the number of possible schedulers, but they also imposes an "heritage" of probabilistic description from execution times to release times for instance.
- *Proving feasibility intervals* is crucial for these approaches that are often used in industry on top of simulation. As worst-case situations are rare events, then observing them or at least observe those events that do provoke later the appearance of worst-case situations is difficult. By proposing an iterative process of composition between different statistical models [17], we provide the basis to build a solution to this essential problem to prove any probabilistic real-time reasoning based on measurements.
- *Providing representativeness* of a measurement-based estimator is the final proof that a probabilistic worst-case reasoning may receive. Our first negative results [24] indicate that the measurement protocol is tighly connected to the statistical estimator and that both must verified properties of reproducibility in order to contribute to a convergence proof.

3.3. Real-Time Systems Compilation

Participant: Dumitru Potop Butucaru.

In the early days of embedded computing, most software development activities were manual. This is no longer true at the low level, where manual assembly coding has been almost completely replaced with the combined use of so-called “high-level” languages (C, Ada, *etc.*) and the use of compilers. This was made possible by the early adoption of standard interfaces that allowed the definition of economically-viable compilation tools with a large-enough user base. These interfaces include not only the programming languages (C, Ada, *etc.*), but also relatively stable microprocessor instruction set architectures (ISAs) or executable code formats like ELF.

The paradigm shift towards fully automated code generation is far from being completed at the system level, mainly due to the slower introduction of standard interfaces. This also explains why real-time scheduling has historically dedicated much of its research effort to verifying the correctness of very abstract and relatively standard implementation models (the task models). The actual construction of the implementations and the abstraction of these implementations as task models drew comparatively less interest, because they were application-dependent and non-portable.

But today the situation is bound to change. First, automation can no longer be avoided, as the complexity of systems steadily increases in both specification size (number of tasks, processors, etc.) and complexity of the objects involved (parallelized dependent tasks, multiple modes and criticalities, many-cores, *etc.*). Second, fully automated implementation is attainable for industrially significant classes of systems, due to significant advances in the standardization of both specification languages (Simulink, Scade, etc.) and of implementation platforms (ARINC, AUTOSAR, *etc.*).

To allow the automatic implementation of complex embedded systems, we advocate for a *real-time systems compilation* approach that combines aspects of both real-time scheduling – including the AAA methodology – and (classical) compilation. Like a classical compiler such as GCC, a real-time systems compiler should use fast and efficient scheduling and code generation heuristics, to ensure scalability. Similarly, it should provide traceability support under the form of informative error messages enabling an incremental trial-and-error design style, much like that of classical application software. This is more difficult than in a classical compiler, given the complexity of the transformation flow (creation of tasks, allocation, scheduling, synthesis of communication and synchronization code, *etc.*), and requires a full formal integration along the whole flow, including the crucial issue of correct hardware/platform abstraction.

A real-time systems compiler should perform precise, conservative timing accounting along the whole scheduling and code generation flow, allowing it to produce safe and tight real-time guarantees. In particular, resource allocation, timing analysis, and code generation must be tightly integrated to ensure that generated code (including communication and synchronization primitive calls) satisfies the timing hypotheses used for scheduling. More generally, and unlike in classical compilers, the allocation and scheduling algorithms must take into account a variety of non-functional requirements, such as real-time constraints, criticality/partitioning, preemptability, allocation constraints, *etc.* As the accent is put on the respect of requirements (as opposed to optimization of a metric, like in classical compilation), resulting scheduling problems are quite different from those of classical compilation.

We have designed and built a prototype real-time systems compiler, called LoPhT, for statically scheduled real-time systems. Results on industrial case studies are encouraging, hinting not only at the engineering potential of the approach, but also at the scientific research directions it opens.

One key issue here is sound hardware/platform abstraction. To prove that it is possible to reconcile performance with predictability in a fully automatic way, we started in the best possible configuration with industrial relevance: statically-scheduled software running on very predictable (yet realistic) platforms. Already, in this case, platform modeling is more complex than the one of classical compilation [0] or real-time scheduling. [0] The objective is now to move beyond this application class to more dynamic classes of specifications implementations, but without losing too much of the predictability and/or effciency.

Efficiency is also a critical issue in practical systems design, and we must invest more in the design of optimizations that improve the *worst-case* behavior of applications and take into account non-functional requirements in a *multi-objective optimization* perspective, but while remaining in the class of low-complexity heuristics to ensure scalability. Optimizations of classical compilation, such as loop unrolling, retiming, and inlining, can serve as inspiration.

Ensuring the safety and efficiency of the generated code cannot be done by a single team. Collaborations on the subject will have to cover at least the following subjects: the interaction between real-time scheduling and WCET analysis, the design of predictable hardware and software architectures, programming language support for efficient compilation, and formally proving the correctness of the compiler.

[0]Because safe timing accounting is needed.

[0]The compiler must perform safe timing accounting, and not rely on experience-derived margins.

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.

GALLIUM Project-Team

3. Research Program

3.1. Programming languages: design, formalization, implementation

Like all languages, programming languages are the media by which thoughts (software designs) are communicated (development), acted upon (program execution), and reasoned upon (validation). The choice of adequate programming languages has a tremendous impact on software quality. By "adequate", we mean in particular the following four aspects of programming languages:

- **Safety.** The programming language must not expose error-prone low-level operations (explicit memory deallocation, unchecked array access, etc) to programmers. Further, it should provide constructs for describing data structures, inserting assertions, and expressing invariants within programs. The consistency of these declarations and assertions should be verified through compile-time verification (e.g. static type-checking) and run-time checks.
- **Expressiveness.** A programming language should manipulate as directly as possible the concepts and entities of the application domain. In particular, complex, manual encodings of domain notions into programmatic notations should be avoided as much as possible. A typical example of a language feature that increases expressiveness is pattern matching for examination of structured data (as in symbolic programming) and of semi-structured data (as in XML processing). Carried to the extreme, the search for expressiveness leads to domain-specific languages, customized for a specific application area.
- **Modularity and compositionality.** The complexity of large software systems makes it impossible to design and develop them as one, monolithic program. Software decomposition (into semi-independent components) and software composition (of existing or independently-developed components) are therefore crucial. Again, this modular approach can be applied to any programming language, given sufficient fortitude by the programmers, but is much facilitated by adequate linguistic support. In particular, reflecting notions of modularity and software components in the programming language enables compile-time checking of correctness conditions such as type correctness at component boundaries.
- **Formal semantics.** A programming language should fully and formally specify the behaviours of programs using mathematical semantics, as opposed to informal, natural-language specifications. Such a formal semantics is required in order to apply formal methods (program proof, model checking) to programs.

Our research work in language design and implementation centers on the statically-typed functional programming paradigm, which scores high on safety, expressiveness and formal semantics, complemented with full imperative features and objects for additional expressiveness, and modules and classes for compositionality. The OCaml language and system embodies many of our earlier results in this area [37]. Through collaborations, we also gained experience with several domain-specific languages based on a functional core, including distributed programming (JoCaml), XML processing (XDuce, CDuce), reactive functional programming, and hardware modeling.

3.2. Type systems

Type systems [39] are a very effective way to improve programming language reliability. By grouping the data manipulated by the program into classes called types, and ensuring that operations are never applied to types over which they are not defined (e.g. accessing an integer as if it were an array, or calling a string as if it were a function), a tremendous number of programming errors can be detected and avoided, ranging from the trivial (misspelled identifier) to the fairly subtle (violation of data structure invariants). These restrictions are also very effective at thwarting basic attacks on security vulnerabilities such as buffer overflows.

The enforcement of such typing restrictions is called type-checking, and can be performed either dynamically (through run-time type tests) or statically (at compile-time, through static program analysis). We favor static type-checking, as it catches bugs earlier and even in rarely-executed parts of the program, but note that not all type constraints can be checked statically if static type-checking is to remain decidable (i.e. not degenerate into full program proof). Therefore, all typed languages combine static and dynamic type-checking in various proportions.

Static type-checking amounts to an automatic proof of partial correctness of the programs that pass the compiler. The two key words here are *partial*, since only type safety guarantees are established, not full correctness; and *automatic*, since the proof is performed entirely by machine, without manual assistance from the programmer (beyond a few, easy type declarations in the source). Static type-checking can therefore be viewed as the poor man's formal methods: the guarantees it gives are much weaker than full formal verification, but it is much more acceptable to the general population of programmers.

3.2.1. Type systems and language design.

Unlike most other uses of static program analysis, static type-checking rejects programs that it cannot prove safe. Consequently, the type system is an integral part of the language design, as it determines which programs are acceptable and which are not. Modern typed languages go one step further: most of the language design is determined by the *type structure* (type algebra and typing rules) of the language and intended application area. This is apparent, for instance, in the XDuce and CDuce domain-specific languages for XML transformations [35], [32], whose design is driven by the idea of regular expression types that enforce DTDs at compile-time. For this reason, research on type systems – their design, their proof of semantic correctness (type safety), the development and proof of associated type-checking and inference algorithms – plays a large and central role in the field of programming language research, as evidenced by the huge number of type systems papers in conferences such as Principles of Programming Languages.

3.2.2. Polymorphism in type systems.

There exists a fundamental tension in the field of type systems that drives much of the research in this area. On the one hand, the desire to catch as many programming errors as possible leads to type systems that reject more programs, by enforcing fine distinctions between related data structures (say, sorted arrays and general arrays). The downside is that code reuse becomes harder: conceptually identical operations must be implemented several times (say, copying a general array and a sorted array). On the other hand, the desire to support code reuse and to increase expressiveness leads to type systems that accept more programs, by assigning a common type to broadly similar objects (for instance, the `Object` type of all class instances in Java). The downside is a loss of precision in static typing, requiring more dynamic type checks (downcasts in Java) and catching fewer bugs at compile-time.

Polymorphic type systems offer a way out of this dilemma by combining precise, descriptive types (to catch more errors statically) with the ability to abstract over their differences in pieces of reusable, generic code that is concerned only with their commonalities. The paradigmatic example is parametric polymorphism, which is at the heart of all typed functional programming languages. Many forms of polymorphic typing have been studied since then. Taking examples from our group, the work of Rémy, Vouillon and Garrigue on row polymorphism [42], integrated in OCaml, extended the benefits of this approach (reusable code with no loss of typing precision) to object-oriented programming, extensible records and extensible variants. Another example is the work by Pottier on subtype polymorphism, using a constraint-based formulation of the type system [40]. Finally, the notion of "coercion polymorphism" proposed by Cretin and Rémy[5] combines and generalizes both parametric and subtyping polymorphism.

3.2.3. Type inference.

Another crucial issue in type systems research is the issue of type inference: how many type annotations must be provided by the programmer, and how many can be inferred (reconstructed) automatically by the type-checker? Too many annotations make the language more verbose and bother the programmer with unnecessary details. Too few annotations make type-checking undecidable, possibly requiring heuristics,

which is unsatisfactory. OCaml requires explicit type information at data type declarations and at component interfaces, but infers all other types.

In order to be predictable, a type inference algorithm must be complete. That is, it must not find *one*, but *all* ways of filling in the missing type annotations to form an explicitly typed program. This task is made easier when all possible solutions to a type inference problem are *instances* of a single, *principal* solution.

Maybe surprisingly, the strong requirements – such as the existence of principal types – that are imposed on type systems by the desire to perform type inference sometimes lead to better designs. An illustration of this is row variables. The development of row variables was prompted by type inference for operations on records. Indeed, previous approaches were based on subtyping and did not easily support type inference. Row variables have proved simpler than structural subtyping and more adequate for type-checking record update, record extension, and objects.

Type inference encourages abstraction and code reuse. A programmer's understanding of his own program is often initially limited to a particular context, where types are more specific than strictly required. Type inference can reveal the additional generality, which allows making the code more abstract and thus more reuseable.

3.3. Compilation

Compilation is the automatic translation of high-level programming languages, understandable by humans, to lower-level languages, often executable directly by hardware. It is an essential step in the efficient execution, and therefore in the adoption, of high-level languages. Compilation is at the interface between programming languages and computer architecture, and because of this position has had considerable influence on the design of both. Compilers have also attracted considerable research interest as the oldest instance of symbolic processing on computers.

Compilation has been the topic of much research work in the last 40 years, focusing mostly on high-performance execution ("optimization") of low-level languages such as Fortran and C. Two major results came out of these efforts: one is a superb body of performance optimization algorithms, techniques and methodologies; the other is the whole field of static program analysis, which now serves not only to increase performance but also to increase reliability, through automatic detection of bugs and establishment of safety properties. The work on compilation carried out in the Gallium group focuses on a less investigated topic: compiler certification.

3.3.1. Formal verification of compiler correctness.

While the algorithmic aspects of compilation (termination and complexity) have been well studied, its semantic correctness – the fact that the compiler preserves the meaning of programs – is generally taken for granted. In other terms, the correctness of compilers is generally established only through testing. This is adequate for compiling low-assurance software, themselves validated only by testing: what is tested is the executable code produced by the compiler, therefore compiler bugs are detected along with application bugs. This is not adequate for high-assurance, critical software which must be validated using formal methods: what is formally verified is the source code of the application; bugs in the compiler used to turn the source into the final executable can invalidate the guarantees so painfully obtained by formal verification of the source.

To establish strong guarantees that the compiler can be trusted not to change the behavior of the program, it is necessary to apply formal methods to the compiler itself. Several approaches in this direction have been investigated, including translation validation, proof-carrying code, and type-preserving compilation. The approach that we currently investigate, called *compiler verification*, applies program proof techniques to the compiler itself, seen as a program in particular, and use a theorem prover (the Coq system) to prove that the generated code is observationally equivalent to the source code. Besides its potential impact on the critical software industry, this line of work is also scientifically fertile: it improves our semantic understanding of compiler intermediate languages, static analyses and code transformations.

3.4. Interface with formal methods

Formal methods collectively refer to the mathematical specification of software or hardware systems and to the verification of these systems against these specifications using computer assistance: model checkers, theorem provers, program analyzers, etc. Despite their costs, formal methods are gaining acceptance in the critical software industry, as they are the only way to reach the required levels of software assurance.

In contrast with several other Inria projects, our research objectives are not fully centered around formal methods. However, our research intersects formal methods in the following two areas, mostly related to program proofs using proof assistants and theorem provers.

3.4.1. Software-proof codesign

The current industrial practice is to write programs first, then formally verify them later, often at huge costs. In contrast, we advocate a codesign approach where the program and its proof of correctness are developed in interaction, and we are interested in developing ways and means to facilitate this approach. One possibility that we currently investigate is to extend functional programming languages such as OCaml with the ability to state logical invariants over data structures and pre- and post-conditions over functions, and interface with automatic or interactive provers to verify that these specifications are satisfied. Another approach that we practice is to start with a proof assistant such as Coq and improve its capabilities for programming directly within Coq.

3.4.2. Mechanized specifications and proofs for programming language components

We emphasize mathematical specifications and proofs of correctness for key language components such as semantics, type systems, type inference algorithms, compilers and static analyzers. These components are getting so large that machine assistance becomes necessary to conduct these mathematical investigations. We have already mentioned using proof assistants to verify compiler correctness. We are also interested in using them to specify and reason about semantics and type systems. These efforts are part of a more general research topic that is gaining importance: the formal verification of the tools that participate in the construction and certification of high-assurance software.

OURAGAN 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 [43], [72], [35] 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 [44] have used them to provide some fast sophisticated factoring algorithms. As a consequence, it is natural to include these objects in our research directions.

3.3. 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 [48], [55]; 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 [46]. 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 [49], [51]. 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 [53], [55]. 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 [42], [50], [47].

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 [52]. Then, theoretical discussions around this conjecture led to a paper on a new approach to the Mahler measure of some 2-variables polynomials [54]. 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 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.

[0] for example the first rational knots are listed at https://team.inria.fr/ouragan/knots
[0] https://isotop.gamble.loria.fr

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).

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.

3.3.3. Static Analysis of x10

x10 is an explicit parallel programming language, originally developped by IBM Research. Parallelism is expressed by the `async / finish` construct (a disymetric variant of `fork / join`), and synchronization uses *clocks*, a sophisticated version of barriers. Programs in this language can be analysed at compile time provided their control statements obey the restrictions of of the polyhedral model. The analysis focuses on the extraction of the *happens before* relation of the subject program, and can be used for the detection of races and deadlocks. A first version of this analysis, which did not take clocks into account, was published in 2013. Its extension to clocked programs is a complex problem, which requires the use of a proof assistant, Coq. Work in collaboration with Alain Ketterlin and Eric Violard (Inria Camus) and Tomofumi Yuki (Inria Cairn).

3.3.4. Toward a Polynomial Model

The polyhedral model is a powerful tool for program analysis and verification, autoparallelization, and optimization. However, it can only be applied to a very restricted class of programs : counted loops, affine conditionals and arrays with affine subscripts. The key mathematical result at the bottom of this model is Farkas lemma, which characterizes all affine function non negative on a polyhedron. Recent mathematical results on the *Positiv Stellen Satz* enable a similar characterization for polynomials positive on a semi-algebraic set. Polynomials may be native to the subject code, but also appears as soon as counting is necessary, for instance when a multidimensional array is linearized or when messages are transmitted through a one dimensional channel. Applying the above theorems allows the detection of polynomial dependences and the construction of polynomial schedules, hence the detection of deadlocks. Code generation from a polynomial schedule is the subject of present work. These methods are applied to the language openStream. Work in collaboration with Albert Cohen and Alain Darte (Xilinx).

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 Gentzen [83] 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 [78], then by Howard and de Bruijn at the end of the 60's [95], [114], 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 [75], [76] stemmed out – a formalism that is both a logic and a programming language and that is at the source of the Coq system [113].

3.1.2. Towards the calculus of constructions

The λ-calculus, defined by Church [73], 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 [66].

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 [105].

In 1985, Coquand and Huet [75], [76] in the Formel team of Inria-Rocquencourt explored an alternative approach based on Girard-Reynolds' system F [84], [109]. 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 [77] 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, Paris 7 and MINES.

In the last six 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 [60] 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 September 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 [67], the more advanced Certified Programming with Dependent Types book by Chlipala [72] 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.

3.3. 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 [85] 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 [102] and Reynolds [108] and started to be studied in an abstract way in the 80's by Felleisen *et al* [81], leading to Parigot's $\lambda\mu$-calculus [106], 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 [101] and Krivine's abstract machine for call-by-name evaluation [98], [97]. 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 [82]: 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 [65]) 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 [80]).

Since a few years, the categorification process has reached logic, with the introduction of homotopy type theory. After a preliminary result that had identified categorical structures in type theory [94], 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 [112], and independently by Burroni [71], 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 [99], [104], [86], [87], and beyond: Petri nets [89] and formal proofs of classical and linear logic have been expressed in this framework [88]. 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 [68].

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 work of Anick [63] and Squier [111], [110] in the 1980s: 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], [91], [92], [93].

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 [31]. 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, each of which corresponding 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 [100], [64]). 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 [10], and in homological algebra [90][31].

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, Guénaël Renault, 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, Ivan Bannwarth, 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, 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 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.

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

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, Guénaël Renault, 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.

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 [70]; it has motivated a serious research effort on the formal analysis of cryptographic protocols, starting with [65] 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* [60]. 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 [58], [44], correspondences (including authentication) [59], and observational equivalences [57]. 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 [52], Signal [68], JFK [45], and Web Services Security [56], 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 [62], results which were extended by formal analysis work using a Dolev-Yao style model of the standard [63]. 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, that models typically ignore. This leads to a situation that a protocol may have been proved secure in theory, but 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# [55]. 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* [73], [49], 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 manual proofs. Programs written in F* can be translated to efficient OCaml, F#, or C for execution [71]. The main ongoing use case of F* is building a verified, drop-in replacement for the whole HTTPS stack in Project Everest [53] (a larger collaboration with Microsoft Research). This includes a verified implementation of TLS 1.2 and 1.3 [54].

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 [52] and Signal [68] 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 [61]. 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 [66] 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* [73], [49] 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 [71]. 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 [54] and of the underlying cryptographic primitives [74].

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 [71]). 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 [50], 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 [46], [47]. 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 [64] [64] and TS* [72], 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:

- 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;
- 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 unconditional 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/

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** [119]. 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], [118], [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 [105]. 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 [106].

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... [68], [110]. 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 [91]. 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 [94] 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], [121], those based on the differentiation of nonlinear flows such as the **return method** [99], [98], and those exploiting the **differential flatness** of the system [103].

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 [66] or shape optimization [74]. Examples of the second type are inactivation principles in human motricity [77] or neurogeometrical models for image representation of the primary visual cortex in mammals [88].

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 [129], geometric measure theory [70] and hypoelliptic operator theory [80].

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 [97]. CAGE's ambition is to be at the core of this development in the years to come.

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. [40] and A. Minca [79]) or mean field interaction models (Garnier-Papanicolaou-Yang [70]).

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 [41], [80], [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 [54], 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 [39] 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 [35].

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 [71], 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 [77]. Thanks to results obtained by Figalli in [63], 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 [55], Duffie, Filipovic and Schachermayer [56]. 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 [62]. A. Alfonsi et al. [33] 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 [53] 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 [36] 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 [24]). 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 [82].

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 [65] 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 [42], 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. 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 [15] 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 [75], 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 [74]). Under a uniform strict hyperbolicity assumption on the characteristic fields, they constructed a multitype version of the sticky particle dynamics. In [76], 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 [69], 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 [83] 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 [72]. 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 [85].

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 [82]). In particular, they characterize the value function as the solution of a reflected BSDE. This property is used in [19] to address American option pricing in markets with imperfections. The Markovian case is treated in [59] 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 [60]. 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 [61]. 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 [84], [85], and relations between information and performance [64].

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 [44]). 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 [48]. An exciting application concerns the number of roots of trigonometric polynomials with random coefficients [49]. 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 [44] are applied in for the approximation of Markov processes (see [52] and [51]). 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 [46], [47], [45]).

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 [81]. 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 [14], 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 [34], 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 [73] 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 [43], 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 [58], [57].

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 ([50]). This method is very general, and A. Alfonsi, A. Kohastu-Higa and M. Hayashi [37] 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 [78] 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 [38] 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.

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 [156] reformulated the problem as a variational problem subject to a convexity constraint. For more general models, and using ideas from Optimal Transportation, Carlier [90] considered the more general c-convexity constraint and proved a general existence result. Using the formulation of [90] McCann, Figalli and Kim [114] gave conditions under which the principal agent problem can be written as an infinite dimensional convex variational problem. The important results of [114] 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 [95], [146], [143].
Goals: So far, the mathematical PA model can be numerically solved for simple utility functions. A Bregman approach inspired by [54] is currently being developed [93] 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* [48], [119] 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 [54], [60], [52] 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 [91] 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) [126] 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 [126]), the porous medium equations, the granular media equation, just to give a few examples. It also finds application in image processing [79]. 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 [97])
- 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 [64]
- 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.

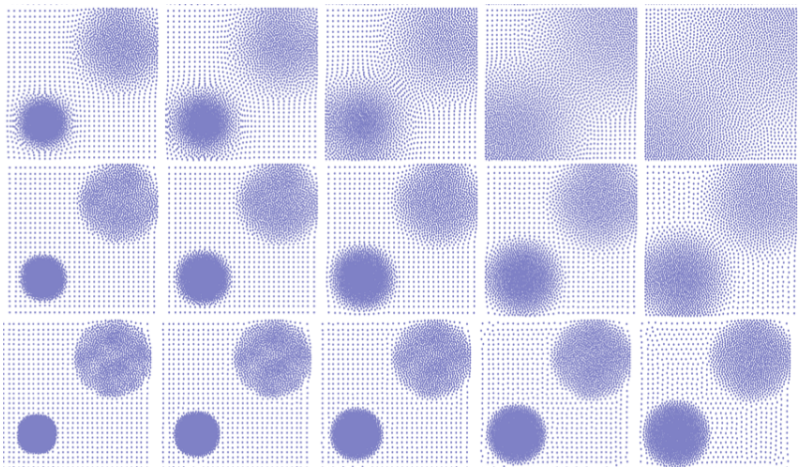

Figure 4. Example of non-linear diffusion equations solved with a JKO flow [56]. The horizontal axis shows the time evolution minimizing the functional $\int \frac{\rho^\alpha}{\alpha-1}$ on the density ρ (discretized here using point clouds, i.e. sum of Diracs' with equal mass). Each row shows a different 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 [94] which leads to very degenerate PDEs [70].
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 [60].
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.

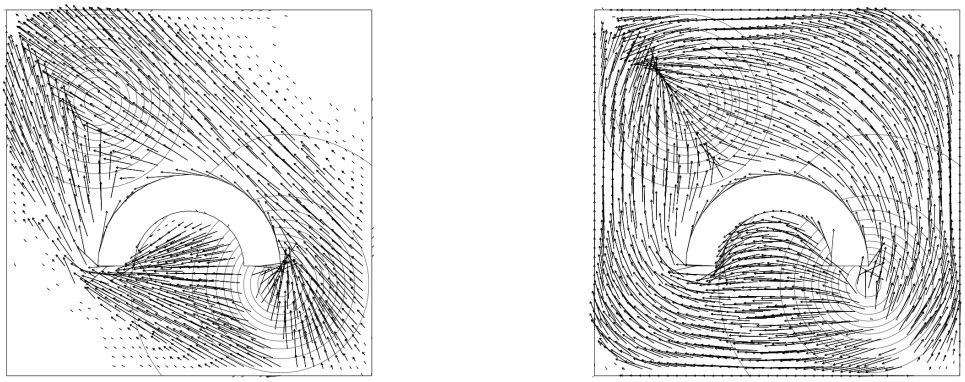

Figure 5. Monge and Wardrop flows of mass around an obstacle [52]. the source/target mass is represented by the level curves. Left : no congestion, Right : congestion.

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 [82] or tracking elastic organ movements [160]. Previous attempts [137], [153] 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 [50] and to define geodesic metrics on the space of shapes and diffeomorphisms [102].
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 [130] with an extremely wide range of potential applications [122]. 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 [53]. 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 [52] and on the analytical side [89] 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 [41].

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 [141] 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 [41], [78].
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 [52], 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) [165], [76], [163]. Our team has already studied flows and geodesics over non-Riemannian shape spaces, which allows for piecewise smooth deformations [102].

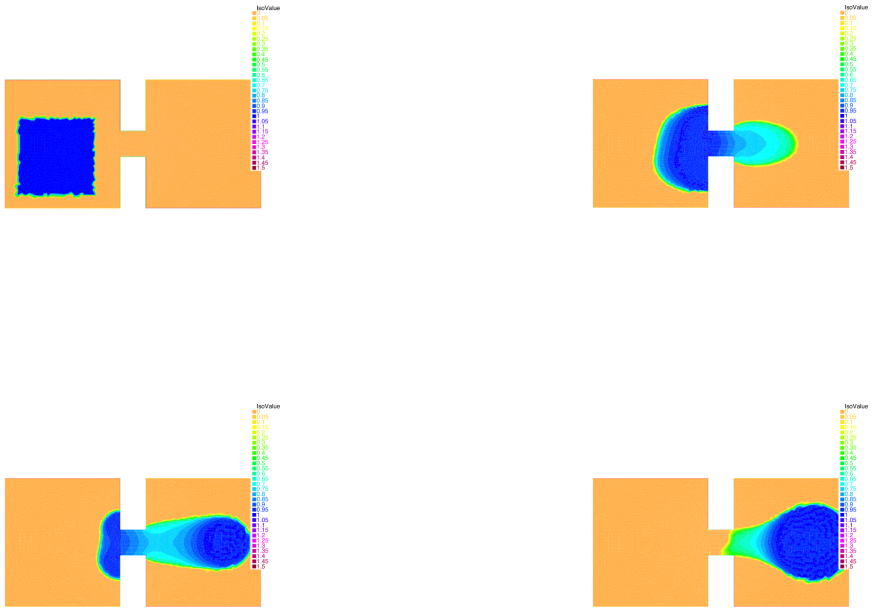

Figure 6. Example of crowd congestion with density dependent speed. The macroscopic density, at 4 different times, of people forced to exit from one room towards a meeting point in a second room.

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 [159], [160] 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 [132], 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 [166] 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 [115]. 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 [115], [116].
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 [164] 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 [65]) for discrete (finite dimensional) problems. We extended this result in a simple infinite dimensional setting, namely sparse regularization of Radon measures for deconvolution [110]. A deep understanding of those continuous inverse problems is crucial to analyze the behavior of their discrete counterparts, and in [111] 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

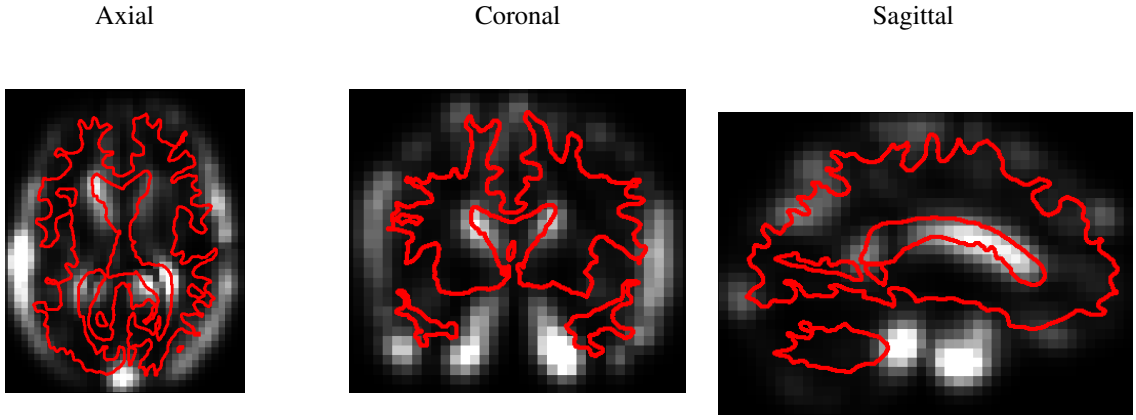

Figure 7. Learning Riemannian metrics in diffeomorphic image matching to capture the brain variability: a diagonal operator that encodes the Riemannian metric is learnt on a template brain out of a collection of brain images. The values of the diagonal operator are shown in greyscale. The red curves represent the boundary between white and grey matter. For more details, we refer the reader to [166], which was a first step towards designing effective and robust metric learning algorithms.

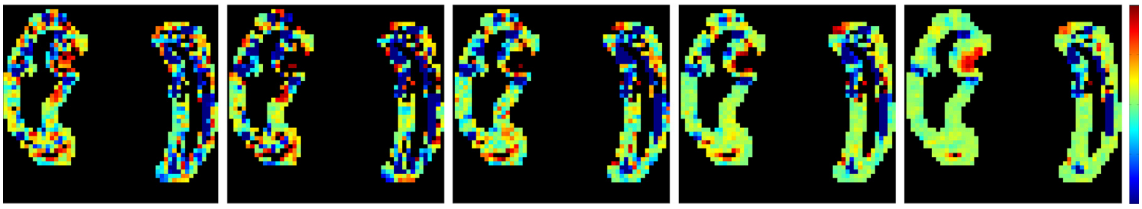

Figure 8. Statistics on initial momenta: In [115], we compared several intersubject transport methodologies to perform statistics on longitudinal evolutions. These longitudinal evolutions are represented by an initial velocity field on the shapes boundaries and these velocity fields are then compared 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 [92], [65] 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 [49], [43], [98]. 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 [110] 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 [152], [112]. This requirement is also at the heart of recently proposed models of cortical processing [151]. 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 [144], [112]. G. Peyré has done several contributions on the definition of geometric wavelets transform and directional texture models, see for instance [152]. Dario Prandi has recently applied methods from sub-Riemannian geometry to image restoration [67].
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 [61]) 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 [123]. 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 [158]. Typical approaches [129] 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 [110].
Goals: We aim at extending the theoretical performance analysis obtained for sparse measures [110] to the set of piecewise regular 2-D and 3-D functions. Some interesting previous work of C. Poon et al [154] (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 [40].

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 [82], [150] 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 [139], which are connected to so-called metamorphoses models in the LDDMM framework [62].
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 [82], [150] 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 [59] for the Monge-Ampère equation and [143] for general non-linear variational problems. Both offer convergence guarantees and are amenable to fast numerical resolution techniques such as Newton solvers. Since [59] 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 [118], a new different and more accurate approach has been proposed by Mirebeau, Benamou and Collino [57]. As shown in [104], 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 [95], Quentin Mérigot [146] and J-M. Mirebeau [143]. 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 [126]) 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 [56] 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 [110], [111] 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 [96].
Our expertise: We have provided the first results on the discrete-to-continous convergence in both sparse regularization variational problems [110], [111] and the static formulation of OT and Wasserstein barycenters [96]
Goals: In a collaboration with Jérôme Malick (Inria Grenoble), our first goal is to generalize the result of [110] to generic partly-smooth convex regularizers routinely used in imaging science and machine learning, a prototypal example being the nuclear norm (see [164] for a review of this class of functionals). Our second goal is to extend the results of [96] to the novel class of entropic discretization schemes we have proposed [54], 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 [87], [110]. 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 [110] 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 [87]). A possible, but very costly, methodology is to ressort to Lasserre's SDP representation hierarchy [131]. 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 [157]), 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 [50], [147], congested optimal transport [81] and to sparse regularizations (see for instance [155] 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 [50], see also [147]. We have also recently developed new proximal schemes that can cope with non-smooth composite objectives functions [155].
Goals: We aim at extending these solvers to a wider class of variational problems, most notably optimization under divergence constraints [52]. 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.

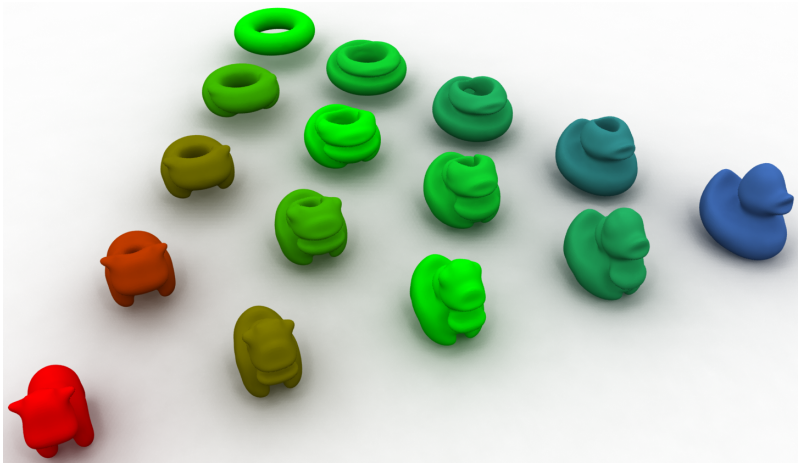

Figure 9. Example of barycenter between shapes computed using optimal transport barycenters of the uniform densities inside the 3 extremal shapes, computed as detailed in [161]. Note that the barycenters are not in general uniform distributions, and we display them as the surface defined by a suitable 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 [108]. As shown in [54], 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 [54] shown how Bregman projections [71] and Dykstra algorithm [46] offer a generic optimization framework to solve a variety of generalized OT problems. Carlier and Dupuis [93] 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 [161]. 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 [101]) 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 [105]) and fluid dynamics (Brenier's weak solutions of the incompressible Euler equation, see [72]). 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).

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 [85], [88]. 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 [62]. 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 [59], [53] 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 [61]. Through a recent experimental work [93], 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 [89] has recently led to a first experimental realization of a full quantum error correcting code improving the coherence time of quantum information [5]. 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.

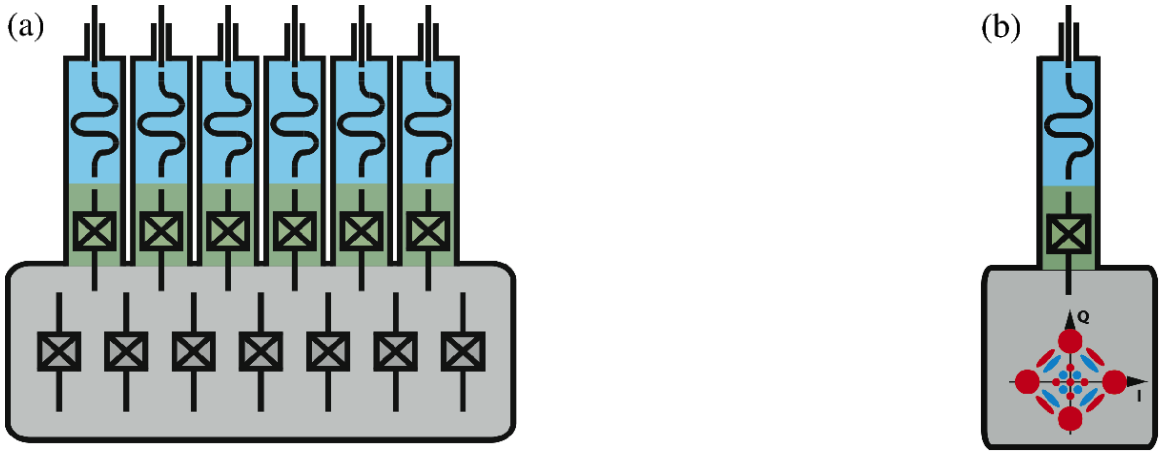

Figure 1. (a) A protected logical qubit consisting of a register of many qubits: here, we see a possible architecture for the Steane code [88] consisting of 7 qubits requiring the measurement of 6 error syndromes. In this sketch, 7 transmon qubits in a high-Q resonator and the measurement of the 6 error 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 [36]. In the context of cavity quantum electro-dynamics (cavity QED) with Rydberg atoms [55], a first experiment on continuous QND measurements of the number of microwave photons was performed by the group at Laboratoire Kastler-Brossel (ENS) [54]. 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 [8] (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 [45], [27], [87], [29]. 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) [44], recent advances in quantum-limited amplifiers [79], [91] have opened doors to high-fidelity non-demolition measurements and real-time feedback for superconducting qubits [56]. 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 [91], [78], [38]. 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 [69].

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 [76] and the closely related coherent feedback [67] 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 [52], single-qubit state stabilization [71], and the creation [31] and stabilization [60], [66][9] 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 [52], [64], [42][12]. The experimental results based on these protocols have illustrated the efficiency of the approach [52][9]. Through these experiments, we exploit the strong dispersive interaction [83] 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 [70], [81], [57][4].

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 [80], [27], [86], [82], [87], [29][7] 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 [29]: 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 [94]; 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 [74]. 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 [56], [38].

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 [32] and is related to quantum trajectories [40], [43]. A modern and mathematical exposure of the diffusive models is given in [30]. In [95] 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 [80], [86]. This stability result is extended to a large class of continuous-time filters in [28]. Further efforts are required to characterize asymptotic and exponential stability. Estimations of convergence rates are available only for quantum non-demolition measurements [33]. Parameter estimations based on measurement data of quantum trajectories can be formulated within such quantum filtering framework [47], [72].

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 [37] that post-selection statistics and "past quantum" state analysis [48] 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 [76], [67]. 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 [7], [9] [52], 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 [75]). 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., [26][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 [7], [9] [52], [70] 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 [51], [50]. We will start with [84] and [77] where, based on a theorem due to Birkhoff [34], 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 [90] and by our recent generalized extension of the latter synchronizing interactions to quantum systems [68].

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 [73], 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 [92], [65] with experimental limitations. The characterization of Kraus maps to stabilize any types of states has also been established [35], 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 [55], [49]. 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 [63].

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 [39] 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 [46] and invariant manifold techniques [41] 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 [58] 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 [52][9] [64].

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.

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.

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 [30], [33], [34]. 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 [31], [32] 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).

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 as a continuation of the BANG project-team, that had been headed by Benoît Perthame from 2003-2013, and in the last years increasingly broadened its subjects as its members developed their own research agendas. 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, to design models and problems under study as well as the related experiments to be carried out by our collaborators in biology or medicine. In this context, our ongoing collaborations with biologists and physicians: the collaboration with St Antoine Hospital in Paris within the Institut Universitaire de Cancérologie of Sorbonne Université (IUC, Luis Almeida, Jean Clairambault, Dirk Drasdo, Alexander Lorz, Benoît Perthame); Institut Jacques Monod (Luis Almeida); the INRA team headed by Human Rezaei and Wei-Feng Xue's team in the university of Canterbury through the ERC Starting Grant SKIPPER^{AD} (Marie Doumic); 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; and the close experimental collaborations that emerged through the former associate team QUANTISS (Dirk Drasdo), particularly at the Leibniz Institute for Working Environment and Human Factors in Dortmund, Germany; or more recently with Yves Dumont at CIRAD, Montpellier, are key points in our project.

Our main objective is the creation, investigation and transfer of new models, methods (for analysis but also for control) and algorithms. 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.

Taking advantage of the last 4-year evaluation of MAMBA (September 2017), we have reorganized the presentation of our research program in three main methodological axes. Two main application axes are presented in the next Section. Evolving along their own logic in close interaction with the methodological axes, they 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 cancer cells, of neurons, of aggregating proteins, etc. whose dynamics are represented by partial differential equations (PDEs), structured in evolving physiological traits, such as cell age, cell size, 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, 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 previous periods of evaluation, many results where obtained in the BANG team on the asymptotic and qualitative behavior of such structured population equations, see e.g. [126], [73], [94], [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), 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 [92], 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 [137]. We investigated the interplay between four processes: nucleation, polymerization, depolymerization and fragmentation.

The ERC Starting Grant SKIPPER^{AD} (Doumic) supported and was the guideline for the study of nucleation, growth and fragmentation equations.

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 [119], [116], [115], [117]. Space may be added as a complementary structure variable provided that something is known of the (Cartesian) geometry of the population [118], which is seldom the case.

Modelling, observation and identification of the spread of infectious diseases

Epidemiological models are made to understand and predict the dynamics of the spread of infectious diseases. We initiated studies with the aim to understand how to use epidemiological data (typically given through incidence rate) in order to estimate the state of the population as well as constants, characteristic of the epidemics such as the transmission rate. The methods rely on observation and identification techniques borrowed from control theory.

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], [103], 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], [112], [144]. 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.

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 [129], followed by [124], [111], [130], [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 [87].

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].

Collaborations

- Nucleation, growth and fragmentation equations: **Juan Calvo**, university of Granada, came for two one-month visits, **Miguel Escobedo**, University of Bilbao (see also Methodological axis 3), **Pierre Gabriel**, University of Versailles-Saint Quentin, former B. Perthame and M. Doumic's Ph.D student, who now co-supervises Hugo Martin's Ph.D thesis. **Piotr Gwiązda**, Polish Academy of Sciences, Poland, **Emil Wiedemann**, University of Bonn, Germany, **Klemens Fellner**, university of Graz, Austria.
- 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, **MAPMO, Orléans**, **Ewelina Zatorska**, University College London, **Anais Khuong**, Francis Crick Institute

3.3. Methodological axis 2: reaction and motion equations for living systems

Personnel

Luis Almeida, Benoît Perthame, Diane Peurichard, Nastassia Pouradier Duteil.

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, REO). 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.

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 [109]. The extension to higher dimensions has been studied in [86]. 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 [102], [108].

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 [140] 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 [110]. 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 [128] for a simple model. This result has motivated the use of another strategy based on viscosity solutions, leading to similar results, in [113].

Since more realistic systems are used in the analysis of medical images, we have extended these studies to include active motion of cells in [127], viscosity in [132] and proved regularity results in [120]. 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 [131], 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 [114].

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*.

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 [97], TRAIL treatment of HeLa cells [74], growth of multicellular spheroids [107], blood detoxification after drug-induced liver damage [139], [101].

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 [133], [96], 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) [13]. 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 [95], a work which inspired also very recently other groups in statistics and probability [75], [105] and was the basis for Adélaïde Olivier's Ph.D thesis [122], [106] and of some of her more recent works [123] (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 [99], was devoted to the stochastic modeling and analysis of protein aggregation, compared both with the deterministic approach traditionally developed in Mamba [137] and with experiments.

Statistical methods decide on subsequently validated mechanism of ammonia detoxification

To identify the mechanisms involved in ammonia detoxification [101], 8 candidate models representing the combination of three possible mechanisms were developed (axis 5). First, the ability of each model to capture the experimental data was assessed by statistically testing the null hypothesis that the data have been generated by the model, leading to exclusion of one of the 8 models. The 7 remaining models were compared among each other by the likelihood ratio. The by far best models were those containing a particular ammonia sink mechanism, later validated experimentally (axis 5). For each of the statistical tests, the corresponding test statistics has been calculated empirically and turned out to be not chi2-distributed in opposition to the usual assumption stressing the importance of calculating the empirical distribution, especially when some parameters are unidentifiable. This year the ammonia detoxification mechanisms have been integrated in a spatial-temporal agent-based model of a liver lobule (the smallest repetitive anatomical unit of liver) and studied for normal and fibrotic liver.

Collaborations

- **Marc Hoffmann,** Université Paris-Dauphine, for the statistical approach to growth and division processes [95], **M. Escobedo**, Bilbao and **M. Tournus**, Marseille, for the deterministic approach.
- **Tom Banks**, North Carolina State University, and **Philippe Moireau,** Inria M3DISIM, for the inverse problem and data assimilation aspects [70], [62]
- **Jan G. Hengstler**, IfADo, Dortmund, Germany

REO Project-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 [46] in the frame of combustion. It was later used to develop the Kiva code [36] at the Los Alamos National Laboratory, or the Hesione code [41], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [39], 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 [42]. 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 [45], and Biot [37] 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 [38]), 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.

3.2. Structure-preserving discretizations 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 requirement is to derive *structure-preserving* methods, i.e., methods that mimic algebraically at the discrete level fundamental properties of the underlying PDEs, such as conservation principles and preservation of invariants. Here, the theoretical questions are closely linked to *differential geometry* and we apply them to the Navier–Stokes equations and to elasto-plasticity. 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 begun this year.

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. This in particular yields 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.

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 shall also prove the numerical optimality of the derived algorithms 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 SkelML to proofs of correctness of numerical algorithms and programs via *mechanical proofs* with Coq.

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
  - Computational kernels for numerical linear algebra
- Validation on numerical simulations

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 http://www.freefem.org/ff++), 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, 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++.

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 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 aproach 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.

DELYS 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 three key challenges:

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.

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 B. Blaszczyszyn and F. Baccelli.
- 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 2017 are summarized in the following sections.

3.2. Distributed network control and smart-grids

Foundation of an entirely new science for distributed control of networks with applications to the stabilization of power grids subject to high volatility of renewable energy production is being developed A. Busic in collaboration with A. Bouillard and Sean Meyn [University of Florida].

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]

3.4. High-dimensional statistical inference for social networks

Community detection and non-regular ramanujan graphs sole a conjecture on the optimality of non-backtracking spectral algorithm for community dectection in sparse stochastic block model graphs, as has been proved by M. Lelarge and L. Massoulié in collaboration with C. Bordenave [IMT Toulouse].

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 standardization working group, 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 will study 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 will study 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 around 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 will investigate self-learning networks. Machine learning approaches, based on experts and forecasters, will be 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 the popularity, expressed by the number of solicitations per day, of 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. Security Trade-off in Constrained Wireless Networks

Ensuring security is a sine qua non condition for the widespread acceptance and adoption of the IoT, in particular in industrial and military applications. While the Public-Key Infrastructure (PKI) approach is ubiquitous on the traditional Internet, constraints in terms of embedded memory, communication bandwidth and computational power make translating PKI to constrained networks non-trivial.

In the IETF 6TiSCH working group, and through the work on Malisa Vucinic as part of the H2020 ARMOUR project, we have started to work on a "Minimal Security" solution at the IETF. This solution is based on pre-shared keying material, and offers mutual authentication between each node in the network and central security authority, replay protection and key rotation.

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 Decompositions

We study new decompositions schemes such as 2-join, skew partitions and others partition problems. These graph decompositions appeared in the structural graph theory and are the basis of some well-known theorems such as the Perfect Graph Theorem. For these decompositions there is a lack of efficient algorithms. We aim at designing algorithms working in $O(nm)$ since we think that this could be a lower bound for these decompositions.

3.1.2. 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 do not need extra memory. We already have 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.

3.1.3. 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 others' 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 from 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 research program is ongoing to carry many important notions of Distributed Computing into a *standard* computational complexity.

3.3. Network Algorithms and Analysis

Based on our scientific foundation on 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 target 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 in 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 the many 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 fast computation time is not guaranteed for all networks. We have already designed such heuristics for diameter computation; understanding the structural properties that enable fast computation time in practice is still an open question.

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. We will 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.

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 [28], [30], [32] or dynamic analysis [51], [53], [59]. 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* [30]. 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* [41] 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 [67]. More recently, more general forms of symbolic execution [28] 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 [25], [26].

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. [35] 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 [29], 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 [17]. 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) [47], 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 [45].

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 [5] [54].

3.1.2.1. Traditional approach.

Using little languages to aid in software development is a tried-and-trusted technique [70] 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 [7]. 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 [6] 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) [3] 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, [18] 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 [36] and its descendant Pharo [22] – or multi-paradigm languages – such as the Scala programming language [56] – 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 [19].

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 [27], [23] 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 [3], 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 [4]. Interestingly, our development crucially relied on an existing Coq formalization, due to Braibant and Pous, [24] 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 [20], [21], [57], [60], [64], [66], [68], [52], [46], 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 [73], 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 [7] to interact with devices, Bossa [6] 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 [3]. 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 [4] is a step in that direction, which belongs to a larger trend of hosting DSLs in dependent type theories [23], [27], [55]. 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 [42], 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.

ALMAnaCH Team

3. Research Program

3.1. Overview and research strands

One of the main challenges in computational linguistics is **to model and to cope with language variation**. Language varies with respect to domain and genre (news wires, scientific literature, poetry, oral transcripts...), sociolinguistic factors (age, background, education; variation attested for instance on social media), geographical factors (dialects) and other dimensions (disabilities, for instance). But language also constantly evolves over all possible time scales. [0] Addressing this variability is still an open issue for NLP. Commonly used approaches, which often rely on supervised and semi-supervised machine learning methods, require huge amounts of annotated data. They are still struggling with the high level of variability found for instance in **user-generated content** or in **non-contemporary texts**.

ALMAnaCH tackles the challenge of language variation in two complementary directions, to which we position a specific activity related to language resources:

3.1.1. Research strand 1

We focus on linguistic representations that are less affected by language variation. It obviously requires us to **stay at a state-of-the-art level in key NLP tasks** such as part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. It also requires improving the **generation of semantic representations (semantic parsing)**. This also involves 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 have to identify, model and take advantage of each available type of contextual information. Addressing these issues enables us to develop new lines of research related to conversational content. Applications include chatbot-based systems and improved information and knowledge extraction algorithms. We especially focus on challenging such specific data sets as domain-specific texts or historical documents, in the larger context of the development of digital humanities.

3.1.2. Research strand 2

Language variation must be better understood and modelled in all its possible realisations. In this regard, we put a strong emphasis on **three 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 with language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus ranging from Old to all forms of Modern French, as well as Indo-European languages in general). In addition, the noise introduced processes such as Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) systems, especially in the context of historical documents, bears similarities with that brought by non-canonical input in user-generated content. This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. [0]

[0] We do not view multilinguality as a case of language variation. Yet multilinguality, a consequence of language diversity, obviously underlies many aspect of ALMAnaCH's research activities.

[0] Other types of language variation could become research topics for ALMAnaCH in the future. This could include dialectal variation (e.g. work on Arabic) as well as the study and exploitation of paraphrases in a broader context than the above-mentioned complexity-based variation.

3.1.3. Research strand 3

Language resource development is not only a technical challenge and a necessary preliminary step to create evaluation data sets for NLP systems as well as training and for machine learning models. It is also a research field in itself, which concerns, among other challenges, (i) the development of semi-automatic and automatic algorithms to speed up the work (e.g. automatic extraction of lexical information, low-resource learning for developing pre-annotation algorithms, transfer methods to leverage tools and/or resources existing for other languages, etc.) and (ii) the development of formal models to represent linguistic information is the best possible way, thus requiring expertise at least both in NLP and in typological and formal linguistics. Language resource development involves the creation of **raw corpora from original sources** as well as the (manual, semi-automatic or automatic) development of **lexical resources** and **annotated corpora**. Such endeavours are domains of expertise of the ALMAnaCH team. This research strand 3 benefits to the whole team and beyond, and both feeds and benefits from the work of the other research strands.

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 (for which cf. Section 7.1). 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. Context-augmented 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 [63], [97] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [93], [90], [96], [21]. 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 [90].

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].

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goals is to explore techniques such as multi-task learning (cf. already [95]) 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.

[0] We ranked 3 for UPOS tagging and 6 for dependency parsing out of 33 participants.
[0] Semantic graph parsing, evaluated on biomedical data, speech and opinion. We ranked 1 in a joint effort with the Stanford NLP team

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 [72]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [62] for document-wise parsing and by [82] 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).

3.2.2. 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.

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.

3.2.3. Chatbots and text generation

Chabots have existed for years (Eliza, Loebner prize). However, they are now becoming the focus of many concrete industrial developments, with the emergence of operational conversational agents and digital assistants (such as Siri). The current approaches mostly rely on the design of scenarios associated with very partial analysis of the requests to fill expected slots and to generate canned answers.

The next generations of such systems will rely on a deeper understanding of the requests, being able to adapt to the specificities of the users, and providing less formatted answers. We believe that chatbots are an interesting and challenging playground to deploy our expertise on knowledge acquisition (to identify concepts and formulations), information extraction based on deeper syntactic representations, context-sensitive analysis (using the thread of exchanges and profile information but also external data sources), and robustness (depending on the possible users' styles).

However, this domain of application also requires working on text generation, starting with simple canned answers and progressively moving to more sophisticated and diverse ones. This work is directly related to another line of research regarding computer-aided text simplification, for which see section 3.3.4 .

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 [60]. 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 [68] 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 [78] 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 [61] and [69]. 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 [94]). 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 [79]. 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 [67], [88], 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 [58]). Lexical simplification, another aspect of text simplification [73], [80], 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.

[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

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 (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 Inriapredecessor 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 [86], [83], [96] 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

[0] From a larger point of view we intend to comply with the so-called FAIR principles (http://force11.org/group/fairgroup/fairprinciples).

languages (see already [98]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

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 [85], [70], [87], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [84], [89]. 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 [70], [87].

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) 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 [92], [81], [91], [76]) 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 [65]. 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.

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., [28], [31], [36], [27], [33]). 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) [30].

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. We conduct work in three areas: weakly supervised and unsupervised algorithms, datasets and benchmarks, and machine intelligence evaluation.

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 [34].

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 [35] 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* [29] , i.e., functional components within a more global cognitive architecture [32]. 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.).

RITS Project-Team

3. Research Program

3.1. Vehicle guidance and autonomous navigation

Participants: Mohammad Abualhoul, Zayed Alsayed, Pierre de Beaucorps, Younes Bouchaala, Pierre Bourre, Raoul de Charette, Carlos Flores, Maximilian Jaritz, Fernando Garrido, Farouk Ghallabi, Shirsendu Halder, Imane Mahtout, Kaouther Messaoud, Francisco Navas, Fawzi Nashashibi, Dinh-Van Nguyen, Renaud Poncelet, Danut-Ovidiu Pop, Luis Roldao, Anne Verroust-Blondet, 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: Zayed Alsayed, Raoul de Charette, Maximilian Jaritz, Farouk Ghallabi, Shirsendu Halder, Kaouther Messaoud, Fawzi Nashashibi, Dinh-Van Nguyen, 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. Cooperative Multi-sensor data fusion

Participant: Fawzi Nashashibi.

Since data are noisy, inaccurate and can also be unreliable or unsynchronized, the use of data fusion techniques is required in order to provide the most accurate situation assessment as possible to perform the perception task. RITS team worked a lot on this problem in the past, but is now focusing on collaborative perception approach. Indeed, the use of vehicle-to-vehicle or vehicle-to-infrastructure communications allows an improved on-board reasoning since the decision is made based on an extended perception.

As a direct consequence of the electronics broadly used for vehicular applications, communication technologies are now being adopted as well. In order to limit injuries and to share safety information, research in driving assistance system is now orientating toward the cooperative domain. Advanced Driver Assistance System (ADAS) and Cybercars applications are moving towards vehicle-infrastructure cooperation. In such scenario, information from vehicle based sensors, roadside based sensors and a priori knowledge is generally combined thanks to wireless communications to build a probabilistic spatio-temporal model of the environment. Depending on the accuracy of such model, very useful applications from driver warning to fully autonomous driving can be performed.

The Collaborative Perception Framework (CPF) is a combined hardware/software approach that permits to see remote information as its own information. Using this approach, a communicant entity can see another remote entity software objects as if it was local, and a sensor object, can see sensor data of others entities as its own sensor data. Last year we developed the basic hardware modules that ensure the well functioning of the embedded architecture including perception sensors, communication devices and processing tools.

Finally, since vehicle localization (ground vehicles) is an important task for intelligent vehicle systems, vehicle cooperation may bring benefits for this task. A new cooperative multi-vehicle localization method using split covariance intersection filter was developed during the year 2012, as well as a cooperative GPS data sharing method.

In the first method, each vehicle estimates its own position using a SLAM (Simultaneous Localization And Mapping) approach. In parallel, it estimates a decomposed group state, which is shared with neighboring vehicles; the estimate of the decomposed group state is updated with both the sensor data of the ego-vehicle and the estimates sent from other vehicles; the covariance intersection filter which yields consistent estimates even facing unknown degree of inter-estimate correlation has been used for data fusion.

In the second GPS data sharing method, a new collaborative localization method is proposed. On the assumption that the distance between two communicative vehicles can be calculated with a good precision, cooperative vehicle are considered as additional satellites into the user position calculation by using iterative methods. In order to limit divergence, some filtering process is proposed: Interacting Multiple Model (IMM) is used to guarantee a greater robustness in the user position estimation.

Accidents between vehicles and pedestrians (including cyclists) often result in fatality or at least serious injury for pedestrians, showing the need of technology to protect vulnerable road users. Vehicles are now equipped with many sensors in order to model their environment, to localize themselves, detect and classify obstacles, etc. They are also equipped with communication devices in order to share the information with other road users and the environment. The goal of this work is to develop a cooperative perception and communication system, which merges information coming from the communications device and obstacle detection module to improve the pedestrian detection, tracking, and hazard alarming.

Pedestrian detection is performed by using a perception architecture made of two sensors: a laser scanner and a CCD camera. The laser scanner provides a first hypothesis on the presence of a pedestrian-like obstacle while the camera performs the real classification of the obstacle in order to identify the pedestrian(s). This is a learning-based technique exploiting adaptive boosting (AdaBoost). Several classifiers were tested and learned in order to determine the best compromise between the nature and the number of classifiers and the accuracy of the classification.

3.1.3. Planning and executing vehicle actions

Participants: Pierre de Beaucorps, Carlos Flores, Fernando Garrido, Imane Mahtout, Fawzi Nashashibi, Francisco Navas, 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). Today, RITS is also investigating the opportunity of fuzzy-based control for specific maneuvers. First results have been recently obtained for reference trajectories following in roundabouts and normal straight roads.

3.2. V2X Communications for cooperative ITS

Participants: Gérard Le Lann, Mohammad Abualhoul, Younes Bouchaala, 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. Visible light and radio communication for cooperative autonomous driving

Participants: Mohammad Abualhoul, Fawzi Nashashibi.

With the extensive development of the automobile industry and the popularity of using personal road vehicles in the last decade, both traffic accidents and road congestion levels have rapidly increased. Taking advantage of advanced wireless communications to enable C-ITS can improve both road fluidity and driver comfort. Ensuring the safety requirements has been the primary interest of the standardization societies dedicated to developing C-ITS applications, in particular with the expected significant demand for a broad range of applications targeting these strict safety requirements. RF communication technology deploying IEEE 802.11p standard for vehicular applications have been dedicated to facilitating relatively medium communication range that supports high data rate for the vehicular environment, where the technology meant to operate within the road safety requirements level.

As a consequence of the accelerated increase of the wireless-based communication devices numbers for ITS applications, the RF communication solutions are pushed toward an insatiable demand for wireless networks data access and a remarkable increase in both latency and channel congestion levels. This instability introduced more usage constraints when C-ITS is required. An example of such applications where the safety requirements and usage constraints might be strictly sharp are the convoy-based ITS applications.

This research effort contributes to the autonomous vehicular communication and urban mobility improvements. The work addresses the main radio-based V2V communication limitations and challenges for ITS hard-safety applications and intends to deploy the vehicular lighting system as a supportive communication solution for convoy-based applications as an IVC [0]-enabled autonomous vehicle. The ultimate objectives of thisresearch was to implement, validate and integrate the VLC system within the existing C-ITS architecture by developing a VLC prototype, together with sufficient hand-over algorithms enabling VLC, RF, and perception-based solutions to ensure the maximum safety requirements and the continuous information exchange between vehicles. The feasibility and efficiency of the VLC-RF system implementation and hand-over algorithms were subjects to perform practical-based in-depth investigations of the system. In addition to the improvement in road capacity by utilizing the convoy-based autonomous driving systems.

[0]Inter Vehicle Communication

3.2.2. 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 [18]. 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.

3.2.3. 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.

3.2.4. Safety-critical communications in intelligent vehicular networks

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 [43]. 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: Mohamed Hadded, Guy Fayolle, Jean-Marc Lasgouttes, Ilias Xydias.

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 [41], 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 [45];
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information [42];
- make the underlying model as sparse as possible, in order to improve BP convergence and quality [44].

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 [39] 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]):

$$Q(x,y)\pi(x,y) = q(x,y)\pi(x) + \widetilde{q}(x,y)\widetilde{\pi}(y) + \pi_0(x,y). \tag{1}$$

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 [40], 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

$$F(x,y,z) = \sum_{i,j,k \geq 0} f(i,j,k) x^i y^j z^k \tag{2}$$

  satisfies the functional equation

$$K(x,y)F(x,y,z) = c(x)F(x,0,z) + \widetilde{c}(y)F(0,y,z) + c_0(x,y), \tag{3}$$

  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.

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 [43]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [76], recursive queries (Datalog), or querying of XML databases [55]. 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 [89] 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 [42], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [51], 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 [81], 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 [59] 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* [43]. 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 [57].

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 [39].

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 [43]. 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* [73] 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 [69], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [70]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [58], [52]. 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 [87] [75], [48].

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 [83], crowdsourcing [50], focused crawling [68], or automatic database tuning [53] critically rely on machine learning techniques, such as classification [72], probabilistic models [67], or reinforcement learning [88].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [78] 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* [85], 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 [84] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [62]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

Having situated Valda's research area within its broader scientific scope, we now move to the discussion of Valda's application domains.

3.2. Research Directions

We now detail three main research axes within the research agenda of Valda. For each axis, we first mention the leading researcher, and other permanent members involved.

3.2.1. Foundations of data management (Luc Segoufin; Serge Abiteboul, Camille Bourgaux, Michaël Thomazo, Pierre Senellart).

Foundations of data management

The systems we are interested in, i.e., for manipulating heterogeneous and confidential data, rapidly changing and massively distributed, are inherently error-prone. The need for formal methods to verify data management systems is best illustrated by the long list of famous leakages of sensitive or personal data that made the front pages of newspapers recently. Moreover, because of the cost in accessing intensional data, it is important to optimize the resources needed for manipulating them.

This creates a need for solid and high-level foundations of DBMS in a manner that is easier to understand, while also facilitating optimization and verification of its critical properties.

In particular these foundations are necessary for various design and reasoning tasks. It allows for clean specifications of key properties of the system such as confidentiality, access control, robustness etc. Once clean specifications are available, it opens the door for formal and runtime verification of the specification. It also permits the design of appropriate query languages – with good expressive power, with limited usage of resources –, the design of good indexes – for optimized evaluation –, and so on. Note that access control policies currently used in database management systems are relatively crude – for example, PostgreSQL offers access control rules on tables, views, or tuples (*row security policies*), but provides no guarantee that these access methods do not contradict each other, or that a user may have access through a query to information that she is not supposed to have access to.

Valda involves leading researchers in the formal verification of data flow in a system manipulating data. Other notable teams involve the WAVE project [0] at U. C. San Diego, and the Business Artifact [0] research program of IBM. One of Valda's objectives is to continue this line of research.

In the short run, we plan to contribute to the state of the art of foundations of systems manipulating data by identifying new scenarios, i.e., specification formalisms, query languages, index structures, query evaluation plans, etc., that allow for any of the tasks mentioned above: formal or runtime verification, optimization etc. Several such scenarios are already known and Valda researchers contributed significantly to their discovery [57], [74], [64], but this research is still in infancy and there is a clear need for more functionalities and more efficiency. This research direction has many facets.

One of the facet is to develop new logical frameworks and new automaton models, with good algorithmic properties (for instance efficient emptiness test, efficient inclusion test and so on), in order to develop a toolbox for reasoning task around systems manipulating data. This toolbox can then be used for higher level tasks such as optimization, verification [57], or query rewriting using views [64].

Another facet is to develop new index structures and new algorithms for efficient query evaluation. For example the enumeration of the output of a query requires the construction of index structures allowing for efficient compressed representation of the output with efficient streaming decompression algorithms as we aim for a constant delay between any two consecutive outputs [82]. We have contributed a lot to this fields by providing several such indexes [74] but there remains a lot to be investigated.

Our medium-term goal is to investigate the borders of feasibility of all the reasoning tasks above. For instance what are the assumptions on data that allow for computable verification 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.

A typical result would be to show that constant delay enumeration of queries is not possible unless the database verify property A and the query property B. Another typical result would be to show that having a robust access control policy verifying at the same time this and that property is not achievable.

[0] http://db.ucsd.edu/WAVE/default.html
[0] http://researcher.watson.ibm.com/researcher/view_group.php?id=2501

Very few such results exist nowadays. If many problems are shown undecidable or decidable, charting the frontier of tractability (say linear time) remains a challenge.

Only when we will have understood the limitation of the method (medium-term goal) 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). This will be our long-term goal.

3.2.2. Uncertainty and provenance of data (Pierre Senellart; Camille Bourgaux, Olivier Cappé, Michaël Thomazo, Luc Segoufin).

Uncertainty and provenance of data

This research axis deals with the modeling and efficient management of data that come with some uncertainty (probabilistic distributions, logical incompleteness, missing values, open-world assumption, etc.) and with provenance information (indicating where the data originates from), as well as with the extraction of uncertainty and provenance annotations from real-world data. Interestingly, the foundations and tools for uncertainty management often rely on provenance annotations. For example, a typical way to compute the probability of query results in probabilistic databases is first to generate the provenance of these query results (in some Boolean framework, e.g., that of Boolean functions or of provenance semirings), and then to compute the probability of the resulting provenance annotation. For this reason, we will deal with uncertainty and provenance in a unified manner.

Valda researchers have carried out seminal work on probabilistic databases [75], [44][7], provenance management [47], incomplete information [46], and uncertainty analysis and propagation in conflicting datasets [65], [41]. These research areas have reached a point where the foundations are well-understood, and where it becomes critical, while continuing developing the theory of uncertain and provenance data management, to move to concrete implementations and applications to real-world use cases.

In the short term, we will focus on implementing techniques from the database theory literature on provenance and uncertainty data management, in the direction of building a full-featured database management add-on that transparently manages provenance and probability annotations for a large class of querying tasks. This work has started recently with the creation of the ProvSQL extension to PostgreSQL, discussed in more details in the following section. To support this development work, we need to resolve the following research question: what representation systems and algorithms to use to support both semiring provenance frameworks [69], extensions to queries with negation [66], aggregation [49], or recursion [80]?

Next, we will study how to add support for incompleteness, probabilities, and provenance annotations in the scenarios identified in the first axis, and how to extract and derive such annotations from real-world datasets and tasks. We will also work on the efficiency of our uncertain data management system, and compare it to other uncertainty management solutions, in the perspective of making it a fully usable system, with little overhead compared to a classical database management system. This requires a careful choice of the provenance representation system used, which should be both compact and amenable to probability computations. We will study practical applications of uncertainty management. As an example, we intend to consider routing in public transport networks, given a probabilistic model on the reliability and schedule uncertainty of different transit routes. The system should be able to provide a user with itinerary to get to have a (probabilistic) guarantee to be at its destination within a given time frame, which may not be the shortest route in the classical sense.

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 [60] on probabilistic query evaluation with the instance-based one of [47], [48].

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.

3.2.3. Personal information management (Serge Abiteboul; Pierre Senellart).

Personal information management

This is a more applied direction of research that will be the context to study issues of interest (see discussion in application domains further).

A typical person today usually has data on several devices and in a number of commercial systems that function as data traps where it is easy to check in information and difficult to remove it or sometimes to simply access it. It is also difficult, sometimes impossible, to control data access by other parties. This situation is unsatisfactory because it requires users to trade privacy against convenience but also, because it limits the value we, as individuals and as a society, can derive from the data. This leads to the concept of Personal Information Management System, in short, a Pims.

A Pims runs, on a user's server, the services selected by the user, storing and processing the user's data. The Pims centralizes the user's personal information. It is a digital home. The Pims is also able to exert control over information that resides in external services (for example, Facebook), and that only gets replicated inside the Pims. See, for instance, [38] for a discussion on the advantages of Pims, as well as issues they raise, e.g. security issues. It is argued there that the main reason for a user to move to Pims is these systems enable great new functionalities.

Valda will study in particular the integration of the user's data. Researchers in the team have already provided important contributions in the context of data integration, notably in the context of the Webdam ERC (2009–2013).

Based on such an integration, Pims can provide a functions, that goes beyond simple query answering:

- Global search over the person's data with a semantic layer using a personal ontology (for example, the data organization the person likes and the person's terminology for data) that helps give meaning to the data;
- Automatic synchronization of data on different devices/systems, and global task sequencing to facilitate interoperating different devices/services;
- Exchange of information and knowledge between "friends" in a truly social way, even if these use different social network platforms, or no platform at all;
- Centralized control point for connected objects, a hub for the Internet of Things; and
- Data analysis/mining over the person's information.

The focus on personal data and these various aspects raise interesting technical challenges that we intend to address.

In the short term, we intend to continue work on the ThymeFlow system to turn it into an easily extendable and deployable platform for the management of personal information – we will in particular encourage students from the M2 *Web Data Management* class taught by Serge and Pierre in the MPRI programme to use this platform in their course projects. The goal is to make it easy to add new functionalities (such as new source *synchronizers* to retrieve data and propagate updates to original data sources, and *enrichers* to add value to existing data) to considerably broaden the scope of the platform and consequently expand its value.

In the medium term, we will continue the work already started that focuses in turning information into knowledge and in knowledge integration. Issues related to intensionality or uncertainty will in particular be considered, relying on the works produced in the other two research axes. We stress, in particular, the importance of minimizing the cost to data access (or, in specific scenarios, the privacy cost associated with obtaining data items) in the context of personal information management: legacy data is often only available through costly APIs, interaction between several Pims may require sharing information within a strict privacy budget, etc. For these reasons, intensionality of data will be a strong focus of the research.

In the long term, we intend to use the knowledge acquired and machine learning techniques to predict the user's behavior and desires, and support new digital assistant functions, providing real *value from data*. We will also look into possibilities for deploying the ThymeFlow platform at a large scale, perhaps in collaboration with industry partners.

WILLOW Project-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.