RESEARCH CENTER
Saclay - Île-de-France

FIELD

Activity Report 2015

Section Scientific Foundations

Edition: 2016-03-21

ALGORITHMICS, PROGRAMMING, SOFTWARE AND ARCHITECTURE

APPLIED MATHEMATICS, COMPUTATION AND SIMULATION

DIGITAL HEALTH, BIOLOGY AND EARTH

NETWORKS, SYSTEMS AND SERVICES, DISTRIBUTED COMPUTING

PERCEPTION, COGNITION AND INTERACTION

COMETE Project-Team

3. Research Program

3.1. Probability and information theory

Participants: Konstantinos Chatzikokolakis, Catuscia Palamidessi, Ehab Elsalamouny, Yusuke Kawamoto, Marco Stronati, Joris Lamare.

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

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

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

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

3.2. Expressiveness of Concurrent Formalisms

Participants: Catuscia Palamidessi, Luis Pino, Frank Valencia.

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

3.3. Concurrent constraint programming

Participants: Michell Guzman, Yamil Salim Perchy, Luis Pino, Frank Valencia.

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

Our research in ccp develops along the following two lines:

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

3.4. Model checking

Participants: Konstantinos Chatzikokolakis, Catuscia Palamidessi.

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

GEOMETRICA Project-Team

3. Research Program

3.1. Mesh Generation and Geometry Processing

Meshes are becoming commonplace in a number of applications ranging from engineering to multimedia through biomedecine and geology. For rendering, the quality of a mesh refers to its approximation properties. For numerical simulation, a mesh is not only required to faithfully approximate the domain of simulation, but also to satisfy size as well as shape constraints. The elaboration of algorithms for automatic mesh generation is a notoriously difficult task as it involves numerous geometric components: Complex data structures and algorithms, surface approximation, robustness as well as scalability issues. The recent trend to reconstruct domain boundaries from measurements adds even further hurdles. Armed with our experience on triangulations and algorithms, and with components from the CGAL library, we aim at devising robust algorithms for 2D, surface, 3D mesh generation as well as anisotropic meshes. Our research in mesh generation primarily focuses on the generation of simplicial meshes, i.e. triangular and tetrahedral meshes. We investigate both greedy approaches based upon Delaunay refinement and filtering, and variational approaches based upon energy functionals and associated minimizers.

The search for new methods and tools to process digital geometry is motivated by the fact that previous attempts to adapt common signal processing methods have led to limited success: Shapes are not just another signal but a new challenge to face due to distinctive properties of complex shapes such as topology, metric, lack of global parameterization, non-uniform sampling and irregular discretization. Our research in geometry processing ranges from surface reconstruction to surface remeshing through curvature estimation, principal component analysis, surface approximation and surface mesh parameterization. Another focus is on the robustness of the algorithms to defect-laden data. This focus stems from the fact that acquired geometric data obtained through measurements or designs are rarely usable directly by downstream applications. This generates bottlenecks, i.e., parts of the processing pipeline which are too labor-intensive or too brittle for practitioners. Beyond reliability and theoretical foundations, our goal is to design methods which are also robust to raw, unprocessed inputs.

3.2. Topological and Geometric Inference

Due to the fast evolution of data acquisition devices and computational power, scientists in many areas are asking for efficient algorithmic tools for analyzing, manipulating and visualizing more and more complex shapes or complex systems from approximative data. Many of the existing algorithmic solutions which come with little theoretical guarantee provide unsatisfactory and/or unpredictable results. Since these algorithms take as input discrete geometric data, it is mandatory to develop concepts that are rich enough to robustly and correctly approximate continuous shapes and their geometric properties by discrete models. Ensuring the correctness of geometric estimations and approximations on discrete data is a sensitive problem in many applications.

Data sets being often represented as point sets in high dimensional spaces, there is a considerable interest in analyzing and processing data in such spaces. Although these point sets usually live in high dimensional spaces, one often expects them to be located around unknown, possibly non linear, low dimensional shapes. These shapes are usually assumed to be smooth submanifolds or more generally compact subsets of the ambient space. It is then desirable to infer topological (dimension, Betti numbers,...) and geometric characteristics (singularities, volume, curvature,...) of these shapes from the data. The hope is that this information will help to better understand the underlying complex systems from which the data are generated. In spite of recent promising results, many problems still remain open and to be addressed, need a tight collaboration between mathematicians and computer scientists. In this context, our goal is to contribute to the development of new mathematically well founded and algorithmically efficient geometric tools for data analysis and processing of complex geometric objects. Our main targeted areas of application include machine learning, data mining, statistical analysis, and sensor networks.

3.3. Data Structures and Robust Geometric Computation

GEOMETRICA has a large expertise of algorithms and data structures for geometric problems.We are pursuing efforts to design efficient algorithms from a theoretical point of view, but we also put efforts in the effective implementation of these results.

In the past years, we made significant contributions to algorithms for computing Delaunay triangulations (which are used by meshes in the above paragraph). We are still working on the practical efficiency of existing algorithms to compute or to exploit classical Euclidean triangulations in 2 and 3 dimensions, but the current focus of our research is more aimed towards extending the triangulation efforts in several new directions of research.

One of these directions is the triangulation of non Euclidean spaces such as periodic or projective spaces, with various potential applications ranging from astronomy to granular material simulation.

Another direction is the triangulation of moving points, with potential applications to fluid dynamics where the points represent some particles of some evolving physical material, and to variational methods devised to optimize point placement for meshing a domain with a high quality elements.

Increasing the dimension of space is also a stimulating direction of research, as triangulating points in medium dimension (say 4 to 15) has potential applications and raises new challenges to trade exponential complexity of the problem in the dimension for the possibility to reach effective and practical results in reasonably small dimensions.

On the complexity analysis side, we pursue efforts to obtain complexity analysis in some practical situations involving randomized or stochastic hypotheses. On the algorithm design side, we are looking for new paradigms to exploit parallelism on modern multicore hardware architectures.

Finally, all this work is done while keeping in mind concerns related to effective implementation of our work, practical efficiency and robustness issues which have become a background task of all different works made by GEOMETRICA.

GRACE Project-Team

3. Research Program

3.1. Algorithmic Number Theory

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

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

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

3.2. Arithmetic Geometry: Curves and their Jacobians

Theme: Arithmetic Geometry: Curves and their Jacobians

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

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

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

(Not every curve is planar—we may have more variables, and more defining equations—but from an algorithmic point of view, we can always reduce to the plane setting.) The *genus*$g_{\mathcal{X}}$ of $\mathcal{X}$ is a non-negative integer classifying the essential geometric complexity of $\mathcal{X}$; it depends on the degree of $F_{\mathcal{X}}$ and on the number of singularities of $\mathcal{X}$. The simplest curves with nontrivial Jacobians are curves of genus 1, known as *elliptic curves*; they are typically defined by equations of the form $y^2 = x^3 + Ax + B$. Elliptic curves are particularly important given their central role in public-key cryptography over the past two decades. Curves of higher genus are important in both cryptography and coding theory.

The curve $\mathcal{X}$ is associated in a functorial way with an algebraic group $J_{\mathcal{X}}$, called the *Jacobian* of $\mathcal{X}$. The group $J_{\mathcal{X}}$ has a geometric structure: its elements correspond to points on a $g_{\mathcal{X}}$-dimensional projective algebraic group variety. Typically, we do not compute with the equations defining this projective variety: there are too many of them, in too many variables, for this to be convenient. Instead, we use fast algorithms based on the representation in terms of classes of formal sums of points on $\mathcal{X}$.

3.3. Curve-Based cryptology

Theme: Curve-Based Cryptology

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

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

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

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

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

3.4. Algebraic Coding Theory

Theme: Coding theory

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

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

These considerations are renewed by the topic of *list decoding* after the breakthrough of Guruswami and Sudan at the end of the nineties. List decoding relaxes the uniqueness requirement of decoding, allowing a small list of candidates to be returned instead of a single codeword. List decoding can reach a capacity close to the Shannon capacity, with zero failure, with small lists, in the adversarial case. The method of

Guruswami and Sudan enabled list decoding of most of the main algebraic codes: Reed–Solomon codes and Algebraic–Geometry (AG) codes and new related constructions “capacity-achieving list decodable codes”. These results open the way to applications again adversarial channels, which correspond to worst case settings in the classical computer science language.

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

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

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

MEXICO Project-Team

3. Research Program

3.1. Concurrency

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar, Serge Haddad, Stefan Schwoon.

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

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

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

3.1.1. Introduction

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

3.1.2. Diagnosis

Participants: Benedikt Bollig, Stefan Haar, Serge Haddad, Stefan Schwoon.

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

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

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

In asynchronous partial-order based diagnosis with Petri nets [63], [64], [68], one unfolds the *labelled product* of a Petri net model $\mathcal{N}$ and an observed alarm pattern $\mathcal{A}$, also in Petri net form. We obtain an acyclic net giving partial order representation of the behaviors compatible with the alarm pattern. A recursive online procedure filters out those runs *(configurations)* that explain *exactly*$\mathcal{A}$. The Petri-net based approach generalizes to dynamically evolving topologies, in dynamical systems modeled by graph grammars, see [46]

3.1.2.1. Observability and Diagnosability

Diagnosis algorithms have to operate in contexts with low observability, i.e., in systems where many events are invisible to the supervisor. Checking *observability* and *diagnosability* for the supervised systems is therefore a crucial and non-trivial task in its own right. Analysis of the relational structure of occurrence nets allows us to check whether the system exhibits sufficient visibility to allow diagnosis. Developing efficient methods for both verification of *diagnosability checking* under concurrency, and the *diagnosis* itself for distributed, composite and asynchronous systems, is an important field for *MExICo*.

3.1.2.2. Distribution

Distributed computation of unfoldings allows one to factor the unfolding of the global system into smaller *local* unfoldings, by local supervisors associated with sub-networks and communicating among each other. In [64], [48], elements of a methodology for distributed computation of unfoldings between several supervisors, underwritten by algebraic properties of the category of Petri nets have been developed. Generalizations, in particular to Graph Grammars, are still do be done.

Computing diagnosis in a distributed way is only one aspect of a much vaster topic, that of *distributed diagnosis* (see [60], [72]). In fact, it involves a more abstract and often indirect reasoning to conclude whether or not some given invisible fault has occurred. Combination of local scenarios is in general not sufficient: the global system may have behaviors that do not reveal themselves as faulty (or, dually, non-faulty) on any local supervisor's domain (compare [45], [51]). Rather, the local diagnosers have to join all *information* that is available to them locally, and then deduce collectively further information from the combination of their views. In particular, even the *absence* of fault evidence on all peers may allow to deduce fault occurrence jointly, see [77], [78]. Automatizing such procedures for the supervision and management of distributed and locally monitored asynchronous systems is a long-term goal to which *MExICo* hopes to contribute.

3.1.3. Contextual nets

Participant: Stefan Schwoon.

Assuring the correctness of concurrent systems is notoriously difficult due to the many unforeseeable ways in which the components may interact and the resulting state-space explosion. A well-established approach to alleviate this problem is to model concurrent systems as Petri nets and analyse their unfoldings, essentially an acyclic version of the Petri net whose simpler structure permits easier analysis [62].

However, Petri nets are inadequate to model concurrent read accesses to the same resource. Such situations often arise naturally, for instance in concurrent databases or in asynchronous circuits. The encoding tricks typically used to model these cases in Petri nets make the unfolding technique inefficient. Contextual nets, which explicitly do model concurrent read accesses, address this problem. Their accurate representation of concurrency makes contextual unfoldings up to exponentially smaller in certain situations. An abstract algorithm for contextual unfoldings was first given in [47]. In recent work, we further studied this subject from a theoretical and practical perspective, allowing us to develop concrete, efficient data structures and algorithms and a tool (Cunf) that improves upon existing state of the art. This work led to the PhD thesis of César Rodríguez in 2014 .

Contexutal unfoldings deal well with two sources of state-space explosion: concurrency and shared resources. Recently, we proposed an improved data structure, called *contextual merged processes* (CMP) to deal with a third source of state-space explosion, i.e. sequences of choices. The work on CMP [79] is currently at an abstract level. In the short term, we want to put this work into practice, requiring some theoretical groundwork, as well as programming and experimentation.

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

3.1.4. Verification of Concurrent Recursive Programs

Participants: Benedikt Bollig, Paul Gastin, Stefan Schwoon.

3.1.5. Dynamic and parameterized concurrent systems

Participants: Benedikt Bollig, Paul Gastin.

In the past few years, our research has focused on concurrent systems where the architecture, which provides a set of processes and links between them, is *static* and *fixed in advance*. However, the assumption that the set of processes is fixed somehow seems to hinder the application of formal methods in practice. It is not appropriate in areas such as mobile computing or ad-hoc networks. In concurrent programming, it is actually perfectly natural to design a program, and claim its correctness, independently of the number of processes that participate in its execution. There are, essentially, two kinds of systems that fall into this category. When the process architecture is static but unknown, it is a parameter of the system; we then call a system *parameterized*. When, on the other hand, the process architecure is generated at runtime (i.e., process creation is a communication primitive), we say that a system is *dynamic*. Though parameterized and dynamic systems have received increasing interest in recent years, there is, by now, no canonical approach to modeling and verifying such systems. Our research program aims at the development of *a theory of parameterized and dynamic concurrent systems*. More precisely, our goal is a *unifying* theory that lays algebraic, logical, and automata-theoretic foundations to support and facilitate the study of parameterized and dynamic concurrent systems. Such theories indeed exist in non-parameterized settings where the number of processes and the way they are connected are fixed in advance. However, parameterized and dynamic systems lack such foundations and often restict to very particular models with specialized verification techniques.

3.1.6. Testing

Participants: Benedikt Bollig, Paul Gastin, Stefan Haar.

3.1.6.1. Introduction

The gap between specification and implementation is at the heart of research on formal testing. The general *conformance testing problem* can be defined as follows: Does an implementation $\mathcal{M}'$ conform a given specification $\mathcal{M}$? Here, both $\mathcal{M}$ and $\mathcal{M}'$ are assumed to have input and output channels. The formal model $\mathcal{M}$ of the specification is entirely known and can be used for analysis. On the other hand, the implementation $\mathcal{M}'$ is unknown but interacts with the environment through observable input and output channels. So the behavior of $\mathcal{M}'$ is partially controlled by input streams, and partially observable via output streams. The Testing problem consists in computing, from the knowledge of $\mathcal{M}$, *input streams* for $\mathcal{M}'$ such that observation of the resulting output streams from $\mathcal{M}'$ allows to determine whether $\mathcal{M}'$ conforms to $\mathcal{M}$ as intended.

In this project, we focus on distributed or asynchronous versions of the conformance testing problem. There are two main difficulties. First, due to the distributed nature of the system, it may not be possible to have a unique global observer for the outcome of a test. Hence, we may need to use *local* observers which will record only *partial views* of the execution. Due to this, it is difficult or even impossible to reconstruct a coherent global execution. The second difficulty is the lack of global synchronization in distributed asynchronous systems. Up to now, models were described with I/O automata having a centralized control, hence inducing global synchronizations.

3.1.6.2. Asynchronous Testing

Since 2006 and in particular during his sabbatical stay at the University of Ottawa, Stefan Haar has been working with Guy-Vincent Jourdan and Gregor v. Bochmann of UOttawa and Claude Jard of IRISA on asynchronous testing. In the synchronous (sequential) approach, the model is described by an I/O automaton with a centralized control and transitions labeled with individual input or output actions. This approach has known limitations when inputs and outputs are distributed over remote sites, a feature that is characteristic of , e.g., web computing. To account for concurrency in the system, they have developed in [70], [52] asynchronous conformance testing for automata with transitions labeled with (finite) partial orders of I/O. Intuitively, this is a “big step” semantics where each step allows concurrency but the system is synchronized before the next big step. This is already an important improvement on the synchronous setting. The non-trivial challenge is now to cope with fully asynchronous specifications using models with decentralized control such as Petri nets.

3.1.6.3. Near Future

Completion of asynchronous testing in the setting without any big-step synchronization, and an improved understanding of the relations and possible interconnections between local (i.e. distributed) and asynchronous (centralized) testing. This has been the objective of the *TECSTES* project (2011-2014), funded by a DIGITEO *DIM/LSC* grant, and which involved Hernán Ponce de Léon and Stefan Haar of *MExICo*, and Delphine Longuet at LRI, University Paris-Sud/Orsay. We have extended several well known conformance (ioco style) relations for sequential models to models that can handle concurrency (labeled event structures). Two semantics (interleaving and partial order) were presented for every relation. With the interleaving semantics, the relations we obtained boil down to the same relations defined for labeled transition systems, since they focus on sequences of actions. The only advantage of using labeled event structures as a specification formalism for testing remains in the conciseness of the concurrent model with respect to a sequential one. As far as testing is concerned, the benefit is low since every interleaving has to be tested. By contrast, under the partial order semantics, the relations we obtain allow to distinguish explicitly implementations where concurrent actions are implemented concurrently, from those where they are interleaved, i.e. implemented sequentially. Therefore, these relations will be of interest when designing distributed systems, since the natural concurrency between actions that are performed in parallel by different processes can be taken into account. In particular, the fact of being unable to control or observe the order between actions taking place on different processes will not be considered as an impediment for testing. We have developped a complete testing framework for concurrent systems, which included the notions of test suites and test cases. We studied what kind of systems are testable in such a framework, and we have proposed sufficient conditions for obtaining a complete test suite as well as an algorithm to construct a test suite with such properties.

A mid-to long term goal (which may or may not be addressed by *MExICo* depending on the availability of staff for this subject) is the comprehensive formalization of testing and testability in asynchronous systems with distributed architecture and test protocols.

3.2. Interaction

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar, Serge Haddad.

3.2.1. Introduction

Systems and services exhibit non-trivial *interaction* between specialized and heterogeneous components. This interplay is challenging for several reasons. On one hand, a coordinated interplay of several components is required, though each has only a limited, partial view of the system’s configuration. We refer to this problem as *distributed synthesis* or *distributed control*. An aggravating factor is that the structure of a component might be semi-transparent, which requires a form of *grey box management*.

Interaction, one of the main characteristics of systems under consideration, often involves an environment that is not under the control of cooperating services. To achieve a common goal, the services need to agree upon a strategy that allows them to react appropriately regardless of the interactions with the environment. Clearly, the notions of opponents and strategies fall within *game theory*, which is naturally one of our main tools in exploring interaction. We will apply to our problems techniques and results developed in the domains of distributed games and of games with partial information. We will consider also new problems on games that arise from our applications.

3.2.2. Distributed Control

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar.

Program synthesis, as introduced by Church [59] aims at deriving directly an implementation from a specification, allowing the implementation to be correct by design. When the implementation is already at hand but choices remain to be resolved at run time then the problem becomes controller synthesis. Both program and controller synthesis have been extensively studied for sequential systems. In a distributed setting, we need to synthesize a distributed program or distributed controllers that interact locally with the system components. The main difficulty comes from the fact that the local controllers/programs have only a partial view of the entire system. This is also an old problem largely considered undecidable in most settings [76], [71], [74], [65], [67].

Actually, the main undecidability sources come from the fact that this problem was addressed in a synchronous setting using global runs viewed as sequences. In a truly distributed system where interactions are asynchronous we have recently obtained encouraging decidability results [66], [56]. This is a clear witness where concurrency may be exploited to obtain positive results. It is essential to specify expected properties directly in terms of causality revealed by partial order models of executions (MSCs or Mazurkiewicz traces). We intend to develop this line of research with the ambitious aim to obtain decidability for all natural systems and specifications. More precisely, we will identify natural hypotheses both on the architecture of our distributed system and on the specifications under which the distributed program/controller synthesis problem is decidable. This should open the way to important applications, e.g., for distributed control of embedded systems.

3.2.3. Adaptation and Grey box management

Participants: Stefan Haar, Serge Haddad.

Contrary to mainframe systems or monolithic applications of the past, we are experiencing and using an increasing number of services that are performed not by one provider but rather by the interaction and cooperation of many specialized components. As these components come from different providers, one can no longer assume all of their internal technologies to be known (as it is the case with proprietary technology). Thus, in order to compose e.g. orchestrated services over the web, to determine violations of specifications or contracts, to adapt existing services to new situations etc, one needs to analyze the interaction behavior of *boxes* that are known only through their public interfaces. For their semi-transparent-semi-opaque nature, we shall refer to them as **grey boxes**. While the concrete nature of these boxes can range from vehicles in a highway section to hotel reservation systems, the tasks of *grey box management* have universal features allowing for generalized approaches with formal methods. Two central issues emerge:

- Abstraction: From the designer point of view, there is a need for a trade-off between transparency (no abstraction) in order to integrate the box in different contexts and opacity (full abstraction) for security reasons.
- Adaptation: Since a grey box gives a partial view about the behavior of the component, even if it is not immediately useable in some context, the design of an adaptator is possible. Thus the goal is the synthesis of such an adaptator from a formal specification of the component and the environment.

Our work on direct modeling and handling of "grey boxes" via modal models (see [61]) was halted when Dorsaf El-Hog stopped her PhD work to leave academia, and has not resumed for lack of staff. However, it should be noted that semi-transparent system management in a larger sense remains an active field for the team, witness in particular our work on diagnosis and testing.

3.3. Management of Quantitative Behavior

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar, Serge Haddad.

3.3.1. Introduction

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

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

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

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

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

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

3.3.2. Probabilistic distributed Systems

Participants: Stefan Haar, Serge Haddad, Claudine Picaronny.

3.3.2.1. Non-sequential probabilistic processes

Practical fault diagnosis requires to select explanations of *maximal likelihood*. For partial-order based diagnosis, this leads therefore to the question what the probability of a given partially ordered execution is. In Benveniste et al. [50], [43], we presented a model of stochastic processes, whose trajectories are partially ordered, based on local branching in Petri net unfoldings; an alternative and complementary model based on Markov fields is developed in [69], which takes a different view on the semantics and overcomes the first model’s restrictions on applicability.

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

In another line of research [54] we have studied the likelihood of occurrence of non-sequential runs under random durations in a stochastic Petri net setting. It remains to better understand the properties of the probability measures thus obtained, to relate them with the models in logical time, and exploit them e.g. in *diagnosis*.

3.3.2.2. Distributed Markov Decision Processes

Participant: Serge Haddad.

Distributed systems featuring non-deterministic and probabilistic aspects are usually hard to analyze and, more specifically, to optimize. Furthermore, high complexity theoretical lower bounds have been established for models like partially observed Markovian decision processes and distributed partially observed Markovian decision processes. We believe that these negative results are consequences of the choice of the models rather than the intrinsic complexity of problems to be solved. Thus we plan to introduce new models in which the associated optimization problems can be solved in a more efficient way. More precisely, we start by studying connection protocols weighted by costs and we look for online and offline strategies for optimizing the mean cost to achieve the protocol. We have been cooperating on this subject with the SUMO team at Inria Rennes; in the joint work [44]; there, we strive to synthesize for a given MDP a control so as to guarantee a specific stationary behavior, rather than - as is usually done - so as to maximize some reward.

3.3.3. Large scale probabilistic systems

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

- Restricting the synchronization between the components as in queuing networks allows to express the steady-state distribution of the model by an analytical formula called a product-form [49].
- Some methods that tackle with the combinatory explosion for discrete-event systems can be generalized to stochastic systems using an appropriate theory. For instance symmetry based methods have been generalized to stochastic systems with the help of aggregation theory [58].
- At last simulation, which works as soon as a stochastic operational semantic is defined, has been adapted to perform statistical model checking. Roughly speaking, it consists to produce a confidence interval for the probability that a random path fulfills a formula of some temporal logic [80] .

We want to contribute to these three axes: (1) we are looking for product-forms related to systems where synchronization are more involved (like in Petri nets), see [9]; (2) we want to adapt methods for discrete-event systems that require some theoretical developments in the stochastic framework and, (3) we plan to address some important limitations of statistical model checking like the expressiveness of the associated logic and the handling of rare events.

3.3.4. Real time distributed systems

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

Timed automata are a well-studied formalism to describe reactive systems that come with timing constraints. For modeling distributed real-time systems, networks of timed automata have been considered, where the local clocks of the processes usually evolve at the same rate [73] [55]. It is, however, not always adequate to assume that distributed components of a system obey a global time. Actually, there is generally no reason to assume that different timed systems in the networks refer to the same time or evolve at the same rate. Any component is rather determined by local influences such as temperature and workload.

3.3.4.1. Implementation of Real-Time Concurrent Systems

Participants: Thomas Chatain, Stefan Haar, Serge Haddad.

This was one of the tasks of the ANR ImpRo.

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

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

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

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

3.3.5. Weighted Automata and Weighted Logics

Participants: Benedikt Bollig, Paul Gastin.

Time and probability are only two facets of quantitative phenomena. A generic concept of adding weights to qualitative systems is provided by the theory of weighted automata [42]. They allow one to treat probabilistic or also reward models in a unified framework. Unlike finite automata, which are based on the Boolean semiring, weighted automata build on more general structures such as the natural or real numbers (equipped with the usual addition and multiplication) or the probabilistic semiring. Hence, a weighted automaton associates with any possible behavior a weight beyond the usual Boolean classification of “acceptance” or “non-acceptance”. Automata with weights have produced a well-established theory and come, e.g., with a characterization in terms of rational expressions, which generalizes the famous theorem of Kleene in the unweighted setting. Equipped with a solid theoretical basis, weighted automata finally found their way into numerous application areas such as natural language processing and speech recognition, or digital image compression.

What is still missing in the theory of weighted automata are satisfactory connections with verification-related issues such as (temporal) logic and bisimulation that could lead to a general approach to corresponding satisfiability and model-checking problems. A first step towards a more satisfactory theory of weighted systems was done in [53]. That paper, however, does not give definite answers to all the aforementioned problems. It identifies directions for future research that we will be tackling.

PARSIFAL Project-Team

3. Research Program

3.1. General overview

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

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

- *Proof normalization*, which treats the state of a computation as a proof term and computation as normalization of the proof terms. General reduction principles such as β-reduction or cut-elimination are merely particular forms of proof normalization. Functional programming is based on normalization [71], and normalization in different logics can justify the design of new and different functional programming languages [48].
- *Proof search*, which views the state of a computation as a a structured collection of formulas, known as a *sequent*, and proof search in a suitable sequent calculus as encoding the dynamics of the computation. Logic programming is based on proof search [77], and different proof search strategies can be used to justify the design of new and different logic programming languages [75].

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

An important organizational principle is structural proof theory, that is, the study of proofs as syntactic, algebraic and combinatorial objects. Formal proofs often have equivalences in their syntactic representations, leading to an important research question about *canonicity* in proofs – when are two proofs "essentially the same?" The syntactic equivalences can be used to derive normal forms for proofs that illuminate not only the proofs of a given formula, but also its entire proof search space. The celebrated *focusing* theorem of Andreoli [50] identifies one such normal form for derivations in the sequent calculus that has many important consequences both for search and for computation. The combinatorial structure of proofs can be further explored with the use of *deep inference*; in particular, deep inference allows access to simple and manifestly correct cut-elimination procedures with precise complexity bounds.

Type theory is another important organizational principle, but most popular type systems are generally designed for either search or for normalization. To give some examples, the Coq system [85] that implements the Calculus of Inductive Constructions (CIC) is designed to facilitate the expression of computational features of proofs directly as executable functional programs, but general proof search techniques for Coq are rather primitive. In contrast, the Twelf system [81] that is based on the LF type theory (a subsystem of the CIC), is based on relational specifications in canonical form (*i.e.*, without redexes) for which there are sophisticated automated reasoning systems such as meta-theoretic analysis tools, logic programming engines, and inductive theorem provers. In recent years, there has been a push towards combining search and normalization in the same type-theoretic framework. The Beluga system [82], for example, is an extension of the LF type theory with a purely computational meta-framework where operations on inductively defined LF objects can be expressed as functional programs.

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

3.2. Inductive and co-inductive reasoning

The team has spent a number of years in designing a strong new logic that can be used to reason (inductively and co-inductively) on syntactic expressions containing bindings. This work is based on earlier work by McDowell, Miller, and Tiu [73] [72] [78] [86], and on more recent work by Gacek, Miller, and Nadathur [4] [63]. The Parsifal team, along with our colleagues in Minneapolis, Canberra, Singapore, and Cachen, have been building two tools that exploit the novel features of this logic. These two systems are the following.

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

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

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

3.3. Developing a foundational approach to defining proof evidence

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

In order to provide our framework with the flexibility that this ambitious plan requires, we have based our design on the most recent advances within the theory of proofs. For a number of years, various team members have been contributing to the design and theory of *focused proof systems* [51] [54] [56] [57] [65] [69] [70] and we have adopted such proof systems as the corner stone for our framework.

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

3.4. Deep inference

Deep inference [66], [68] is a novel methodology for presenting deductive systems. Unlike traditional formalisms like the sequent calculus, it allows rewriting of formulas deep inside arbitrary contexts. The new freedom for designing inference rules creates a richer proof theory. For example, for systems using deep inference, we have a greater variety of normal forms for proofs than in sequent calculus or natural deduction systems. Another advantage of deep inference systems is the close relationship to categorical proof theory. Due to the deep inference design one can directly read off the morphism from the derivations. There is no need for a counter-intuitive translation.

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

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

3.5. Proof nets and atomic flows

Proof nets and atomic flows are abstract (graph-like) presentations of proofs such that all "trivial rule permutations" are quotiented away. Ideally the notion of proof net should be independent from any syntactic formalism, but most notions of proof nets proposed in the past were formulated in terms of their relation to the sequent calculus. Consequently we could observe features like “boxes” and explicit “contraction links”. The latter appeared not only in Girard’s proof nets [64] for linear logic but also in Robinson’s proof nets [83] for classical logic. In this kind of proof nets every link in the net corresponds to a rule application in the sequent calculus.

Only recently, due to the rise of deep inference, new kinds of proof nets have been introduced that take the formula trees of the conclusions and add additional “flow-graph” information (see e.g., [6], [5] and [67]. On one side, this gives new insights in the essence of proofs and their normalization. But on the other side, all the known correctness criteria are no longer available.

This directly leads to the following research questions investigated by members of the Parsifal team:

- Finding (for classical logic) a notion of proof nets that is deductive, i.e., can effectively be used for doing proof search. An important property of deductive proof nets must be that the correctness can be checked in linear time. For the classical logic proof nets by Lamarche and Straßburger [6] this takes exponential time (in the size of the net).
- Studying the normalization of proofs in classical logic using atomic flows. Although there is no correctness criterion they allow to simplify the normalization procedure for proofs in deep inference, and additionally allow to get new insights in the complexity of the normalization.

3.6. Cost Models and Abstract Machines for Functional Programs

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

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

Nonetheless, it turns out that the number of β-steps is a reasonable time cost model, i.e.,it is polynomially related to those of Turing machines and RAM. For the special case of *weak evaluation* (i.e., reducing only β-steps that are not under abstractions)—which is used to model functional programming languages—this is a relatively old result due to Blelloch and Greiner [53] (1995). It is only very recently (2014) that the strong case—used in the implementation models of proof assistants—has been solved by Accattoli and Dal Lago [49].

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

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

POSTALE Team

3. Research Program

3.1. Architectures and program optimization

In this research topic, we focus on optimizing resources in a systematic way for the programmer by addressing fundamental issues like optimizing communication and data layout, generating automatically optimized codes via Domain Specific Languages (DSL), and auto-tuning of computer systems.

3.1.1. Optimization techniques for data and energy

3.1.1.1. Scientific context

Among the main challenges encountered in the race towards performance for supercomputers are energy (consumption, power and heat dissipation) and the memory/communication wall. This research topic addresses more specialized code analysis and optimization techniques as well as algorithmic changes in order to meet these two criteria, both from an expert - meaning handmade code transformations - or automatic - meaning compile time or run time - point of view.

Memory/communication wall means that processor elementary clock cycle decreases more rapidly over years than data transfer whether vertically between memory-ies and CPU (memory access) or horizontally between processors (data transfer). Moreover current architectures include complex memory features such as deep memory hierarchies, shared caches between cores, data alignment constraints, distributed memories etc. As a result data communication and data layout are becoming the bottleneck to performance and most program transformations aim at organizing them carefully and possibly avoiding or minimizing them. Energy consumption is also a limitation for today's processor performance. Then the options are either to design processors that consume less energy or, at the software level, to design energy-saving compilers and algorithms.

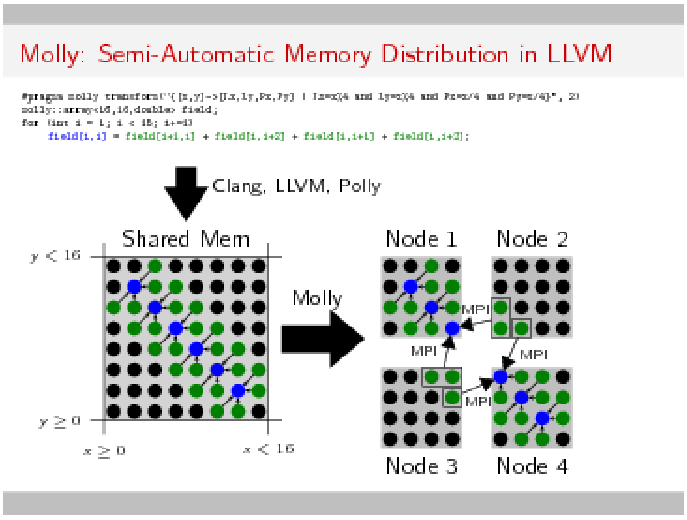
In general, the memory and energy walls are tackled with the same kind of program transformations that consist of avoiding as much as possible data communication [143] but considering these issues separately offers a different perspective. In this research axis, we focus on data/memory and energy/power optimization that include handmade or automatic compiler, code and algorithm optimizations. The resulting tools are expected to be integrated in other Postale topics related to auto-tuning [79], code generation [69] or communication-avoiding algorithms [37], [98].

3.1.1.2. Activity description and recent achievements

3.1.1.2.1. Optimization for data:

Program data transformation - data layout, data transfers. Postale has been addressing these issues in the past ANR PetaQCD project described in [49], [50] and in the PhD thesis of Michael Kruse [99]. The latter describes handmade data layout optimizations for optimizing a 4D stencil computation taking into account the BlueGene Q features. It also presents the Molly software based on the LLVM (Low Level Virtual Machine) Polly optimizing compiler that automatically generates code for MPI data transfers (see Figure 1 that shows an example of code generating a decomposition of a stencil computation into 4 subdomains and how data are exchanged between subdomains).

Data layout is still a critical point that Postale will address. The DSL [69] approach allows us to consider data layout globally, providing then an opportunity to study aggressive layouts without transformation penalty. We will also seize this opportunity to investigate the data layout problem as a new dimension of the CollectiveMind [79] optimization topic.



Molly: Semi-Automatic Memory Distribution in LLVM

```
#pragma molly transform("{ [x,y]->[Lx,Ly,Px,Py] | Lx=x%4 and Ly=y%4 and Px=x/4 and Py=y/4}", 2)
molly::array<16,16,double> field;
for (int i = 1; i < 15; i+=1)
    field[i,i] = field[i+1,i] + field[i,i+2] + field[i,i+1] + field[i,i+2];
```

Figure 1. Automatic generation of subdomains using the Molly software.

Algorithm transformation - automating communication avoiding algorithms. This part is related to the Postale work on numerical algorithms. It originates from a research grant application elaborated with the former PetaQCD [50] team and the Inria Alpine project-team. One essential research direction consists of providing a set of high level optimizations that are generally out of reach from a traditional compiler approach. Among these optimizations, we consider communication-avoiding transformations and address the current open question of integrating these transformations in the polyhedral model in order to make them available in most software environments. Communication-avoiding algorithms improve parallelism and decrease communication requirements by ignoring some of dependency constraints at the frontiers of subdomains. Integrating communication-avoiding transformations is challenging first because these transformations change code semantics, which is unsual in program transformations, second because the validity of these transformations relies on numerical properties of the underlying transformed algorithms. This requires both compiler and algorithm skills since these transformations have important impact on the numerical stability and convergence of algorithms. Tools for the automatic generation of these transformed algorithms have two kinds of application. First, they accelerate the fastidious task of reprogramming for testing numerical properties. They may even be incorporated in an iterative tool for systematically evaluating these properties. Second, if these transformations are formalized we can consider generating different versions on line at run time, to adapt automatically algorithms to run time values [51]. In particular we plan to address s-steps algorithms [119] in iterative methods as these program transformations are similar to loop unrolling and ghosting (inverse of loop peeling). These are aggressive transformations and special preconditioning is needed in order to ensure convergence.

3.1.1.2.2. Optimizing energy:

In this topic there are two main research directions. The first one is about reversible computing based on the Landauer's conjecture that heat dissipation is produced by information erasing. The second one is on actual measurements of energy/power of program execution and on understanding which application features are the most likely to save or consume energy.

Regarding **reversible computing**, the Landauer's hypothesis - still in discussion among physicists - says that erasing one bit of information dissipates energy, independently from hardware. This implies that energy saving algorithms should avoid as much as possible erasing information: it should be possible to recover values of variables at any time in program execution. In a previous work we have analyzed the impact of making computing DAG (Directed Acyclic Graphs) reversible [47]. We have also used reversible computing in register allocation by enabling value rematerialization also by reverse computing [48]. We are now working on characterizing algorithms by the amount of input and output data that have to be added to make algorithms reversible. We also plan to analyze mixed precision numerical algorithms [36] from this perspective.

Another research direction concerns **energy and power profiling and optimizing**. Understanding and monitoring precise energetic behavior of current programs is still a not easy task for the programmer or the compiler. One can measure it with wattmeters, or perform processor simulations or use hardware counters or sensors, or approximate it by the number of data that are communicated [144]. Especially on supercomputers or cloud framework it might be impossible to get this information. Besides making experiments on energy and power profiling [114], this research axis also includes the analysis of programming features that are the key parameters for saving energy. The ultimate goal is to have a cost model that describes the program energetic behavior of programs for the programmer or compiler being able to control it. One obvious key parameter is the count of memory accesses but one can also think of regularity features such as constant strides memory access, whether the code is statically or dynamically controlled, regularity/predictability conditional branchs. We have already performed this kind of analysis in the context of value prediction techniques where we designed entropy based criteria for estimating the predictibility of the sequence of values of some variables [115].

3.1.1.3. Research tracks for the 4 next years

Short term objectives are related to handmade or semi-automatic profiling and optimization of current scientific or image processing challenging applications. This gives a very good insight and expertise over state of the art applications and architectures. This know-how can be exploited under the form of libraries. This includes performance profiling, analysis of the energetic behavior of applications, and finding hot spots and focus optimization on these parts. This also implies to implement new numerical algorithms such as the communication-avoiding algorithms. Mid term objectives are to go forward to the automatization or semi-automatization of these techniques. Long term objectives are to understand the precise relationship between physics and computation both in programs as in reversible computing and in algorithms like in algorithmic thermodynamics [46]. The path is to define a notion of energetic complexity, which we intend to do it with the Galac team at Laboratoire de Recherche en Informatique.

3.1.2. Generative programming for new parallel architectures

3.1.2.1. Scientific context

Design, development and maintenance of high-performance scientific code is becoming one of the main issue of scientific computing. As hardware is becoming more complex and programming tools and models are proposed to satisfy constantly evolving applications, gathering expertise in both any scientific field and parallel programming is a daunting task. The natural conclusion is then to provide software design tools such that non-experts in computer science are able to produce non-trivial yet efficient codes on modern hardware architectures at their disposal. These tools can be divided in two types:

- **Compilers**. Compilers can be designed to either automatically derive parallel version of sequential codes or to support specific annotations to do so. Various successful examples include ISPC [122], SPADE [152] or GCC and its support for polyhedral compilation [125]. By offloading these tasks to compilers, the performance of the resulting codes is free of any overhead and the amount of user input is minimized. However, the scope and applicability of these techniques are fragile and can be hindered by complex code flow, inadequate data types or the use of high level languages features.
- **Libraries**. The unability of compilers to handle complex semantic is often mitigated by the design of libraries. Libraries can expose an arbitrary high level of abstraction through abstract data types and functions operating on them. User code is then expressed as a combination of function calls over instances of these data types. Different level of abstraction for parallel systems are available ranging from linear algebra [28], [95], image processing [56] to graph algorithms [138]. The main limitation of this approach is the lack of inter-procedural optimizations and the inherent divergence in API among vendors and targeted systems.

One emerging solution is to combine aspects of both solutions by designing systems which are able to provide abstraction and performance. One such approach is the design and development of **Domain Specific Languages** (or DSL) and more precisely, **Domain Specific Embedded Languages** (DSEL). DSLs [139] are non-general purpose, declarative language that simplify development by allowing users to express “the problem to solve” instead of “how to solve it”. Actual code generation is then left to a proper compiler, interpreter or code generator that use high-level abstraction analysis and potential knowledge about target hardware to ensure performance. SCALA – and more precisely the FORGE tool [141] – is one of the most successful attempt at applying such techniques to parallel programming. DSELs differ from regular DSLs in the fact that they exist as a subset of an existing general purpose language. Often implemented as **Active Libraries** [151], they perform high-level optimizations based on a semantic analysis of the code before any real compilation process.

3.1.2.2. Activity description and recent achievements

In this research, we investigate the impact and applicability of software design methods based on DSELs to parallel programming and we study the portability and forward scalability of such programs. To do so, we investigate **Generative Programming** [62] applied to parallel programming.

Generative Programming is based on the hypothesis that any complex software system can be split into a list of interchangeable components (with clearly identified tasks) and a series of generators that combine components by following rules derived from an a priori domain specific analysis. In particular, we want to show that integrating the architectural support as another generative component of the set of tools leads to a better performance and an easier development on embedded or custom architecture targets (see Figure 2).

The application of Generative Programming allows us to build active libraries that can be easily re-targeted, optimized and deployed on a large selection of hardware systems. This is done by decoupling the abstract description of the DSEL from the description of hardware systems and the generation of hardware agnostic software components.

Current applications of this methodology include:

- BOOST.SIMD [70] is a C++ library for portable SIMD computations. It uses architecture aware generative programming to generate zero-overhead SIMD code on a large selection of platforms (from SSE to AVX2, Xeon Phi, PowerPC and ARM). Its interface is made so it is totally integrated into modern C++ design strategy based on the use of generic code and calls to the standard template libraries. In most cases, BOOST.SIMD delivers performance on the par with hand written SIMD code or with autovectorizers.
- NT^2 [69], [75] is a C++ library which implements a DSEL similar to MATLAB while providing automatic parallelization on SIMD systems, multicores and GPGPUs. NT^2 uses the high level of abstraction brought by the MATLAB API to detect, analyze and generate efficient loop nests taking care of every level of parallel hardware available. NT^2 eases the design of scientific computing application prototypes while delivering a significant percentage of the peak performance.

Our work uses a methodogy similar to SCALA [120], and more specifically, the DeLITE [142] toolset. Both approach rely on extracting high level, domain specific information from user code to optimize HPC applications. If our approach tries to maximize the use of compile-time optimization, DeLITE uses a runtime approach due to its reliance on the JAVA language.

In terms of libraries, various existing Scientific Computing library in C++ are actually available. The three most used are Armadillo [137], which shares a MATLAB-like API with our work, Blaze [55] which supports a similar cost based system for optimizing code and Eigen [86]. Our main feature compared to these solutions is the fact that hardware support is built-in the library core instead of beign tacked on the existing library, thus allowing us to support a larger amount of hardware.

3.1.2.3. Research tracks for the 4 next years

At short term, research and development on BOOST.SIMD and NT^2 will explore the applicability of our code generation methodology on distributed system, accelerators and heterogeneous systems. Large system support like Blue Gene/Q and other similar super-computer setup has been started.

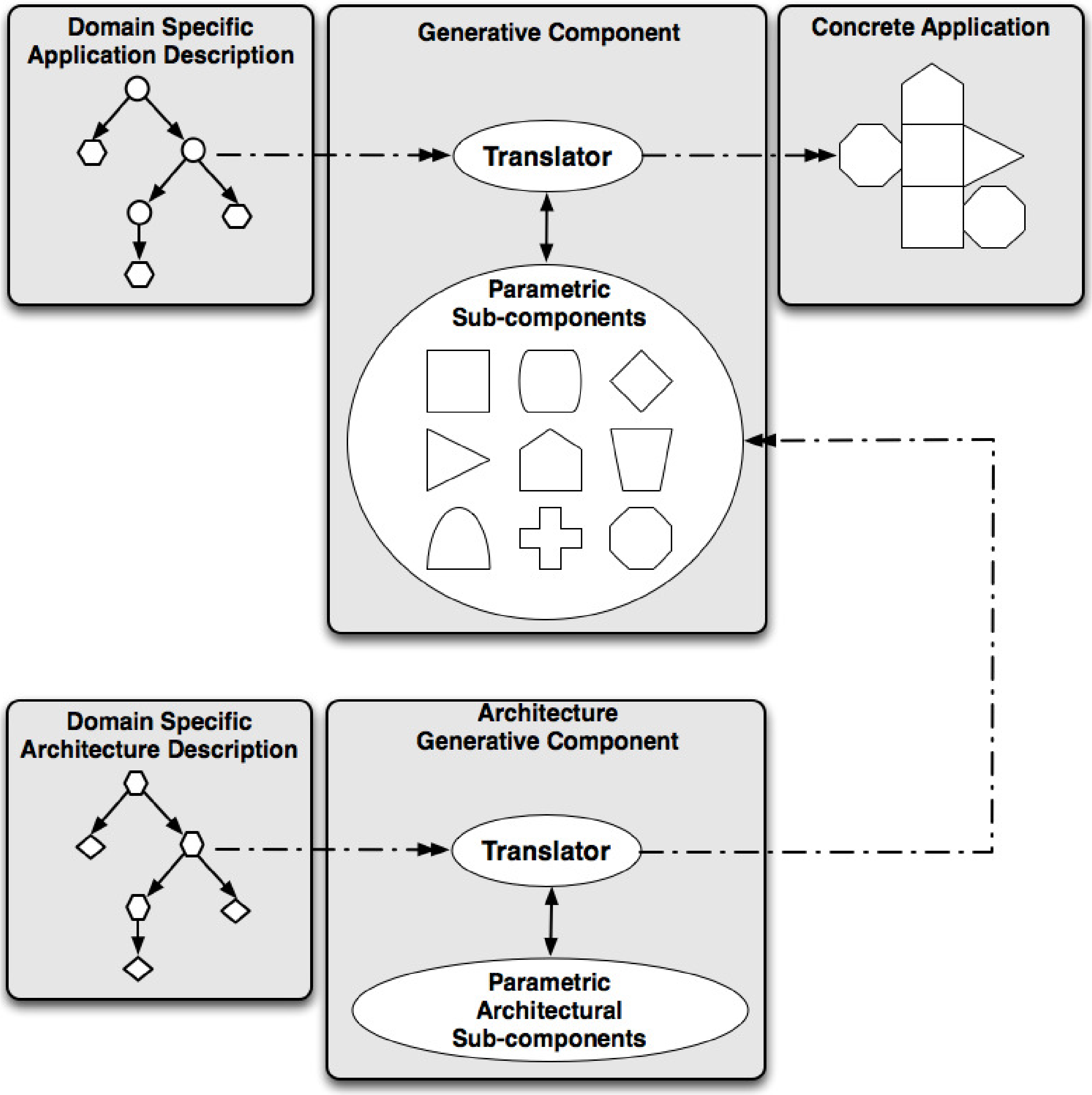


Figure 2. Principles of Architecture Aware Generative Programming

Another axis of research is to apply generative programing to other scientific domain and to propose other domain specific tools using efficient code generators. Such a work has been started to explore the impact of generative programming on the design of portable linear algebra algorithms with an going PhD thesis on automatic generation of linear algebra software.

A mid-term objective is to bridge the gap with the Data Analytics community in order to both extract new expertise on how to make Big Data related issues scalable on modern HPC hardware and to provide tools for Data Analytics practitioners based on this collaboration.

On a larger scope, the implication of our methodology on language design will be explored. First by proposing evolution to C++ (as for example with our SIMD proposal [71]) so that generative programming can become a first class citizen in the language itself. Second by exploring how this methodology can be extended to other languages [85] or to other runtime systems including Cloud computing systems and JIT support. Application to other performance metric like power consumption is also planned [156].

3.1.3. Systematizing and automating program optimization

3.1.3.1. Scientific context

Delivering faster, more power efficient and reliable computer systems is vital for our society to continue innovation in science and technology. However, program optimization and hardware co-design became excessively time consuming, costly and error prone due to an enormous number of available design and optimization choices, and complex interactions between all software and hardware components. Worse, multiple characteristics have to be always balanced at the same time including execution time, power consumption, code size, memory utilization, compilation time, communication costs and reliability using a growing number of incompatible tools and techniques with many ad-hoc and intuition based heuristics. As a result, nearly peak performance of the new systems is often achieved only for a few previously optimized and not necessarily representative benchmarks while leaving most of the real user applications severely underperforming. Therefore, users are often forced to resort to a tedious and often non-systematic optimization of their programs for each new architecture. This, in turn, leads to an enormous waste of time, expensive computing resources and energy, dramatically increases development costs and time-to-market for new products and slows down innovation [27], [25], [32], [66].

3.1.3.2. Activity description and recent achievements

For the european project MILEPOST (2006-2009) [26], we, for the first time to our knowledge, attempted to address above challenges in practice with several academic and industrial partners including IBM, CAPS, ARC (now Synopsys) and the University of Edinburgh by combining automatic program optimization and tuning, machine learning and a public repository of experimental results. As a part of the project, we established a non-profit cTuning association (cTuning.org) that persuaded the community to voluntarily support our open source tools and repository while sharing benchmarks, data sets, tools and machine learning models even after the project. This approach, highly prised by the European Commission, Inria and the international community, helped us to substitute and automatically learn best compiler optimization heuristics by crowdsourcing auto-tuning (processing a large amount of performance statistics or "big data" collected from many users to classify application and build predictive models) [26], [77], [78]. However, it also exposed even more fundamental challenges including:

- Lack of common, large and diverse benchmarks and data sets needed to build statistically meaningful predictive models;
- Lack of common experimental methodology and unified ways to preserve, systematize and share our growing optimization knowledge and research material from the community including benchmarks, data sets, tools, tuning plugins, predictive models and optimization results;
- Problem with continuously changing, “black box” and complex software and hardware stack with many hardwired and hidden optimization choices and heuristics not well suited for auto-tuning and machine learning;
- Difficulty to reproduce performance results from the cTuning.org database submitted by the community due to a lack of full software and hardware dependencies;

- Difficulty to validate related auto-tuning and machine learning techniques from existing publications due to a lack of culture of sharing research artifacts with full experiment specifications along with publications in computer engineering.

As a result, we spent a considerable amount of our "research" time on re-engineering existing tools or developing new ones to support auto-tuning and learning. At the same time, we were trying to somehow assemble large and diverse experimental sets to make our research and experimentation on machine learning and data mining statistically meaningful. We spent even more time when struggling to reproduce existing machine learning-based optimization techniques from numerous publications. Worse, when we were ready to deliver auto-tuning solutions at the end of such tedious developments, experimentation and validation, we were already receiving new versions of compilers, third-party tools, libraries, operating systems and architectures. As a consequence, our developments and results were already potentially outdated even before being released while optimization problems considerably evolved.

We believe that these are major reasons why so many promising research techniques, tools and data sets for auto-tuning and machine learning in computer engineering have a life span of a PhD project, grant funding or publication preparation, and often vanish shortly after. Furthermore, we witness diminishing attractiveness of computer engineering often seen by students as "hacking" rather than systematic science. Many recent long-term research visions acknowledge these problems for computer engineering and many research groups search for "holy grail" auto-tuning solutions but no widely adopted solution has been found yet [25], [66].

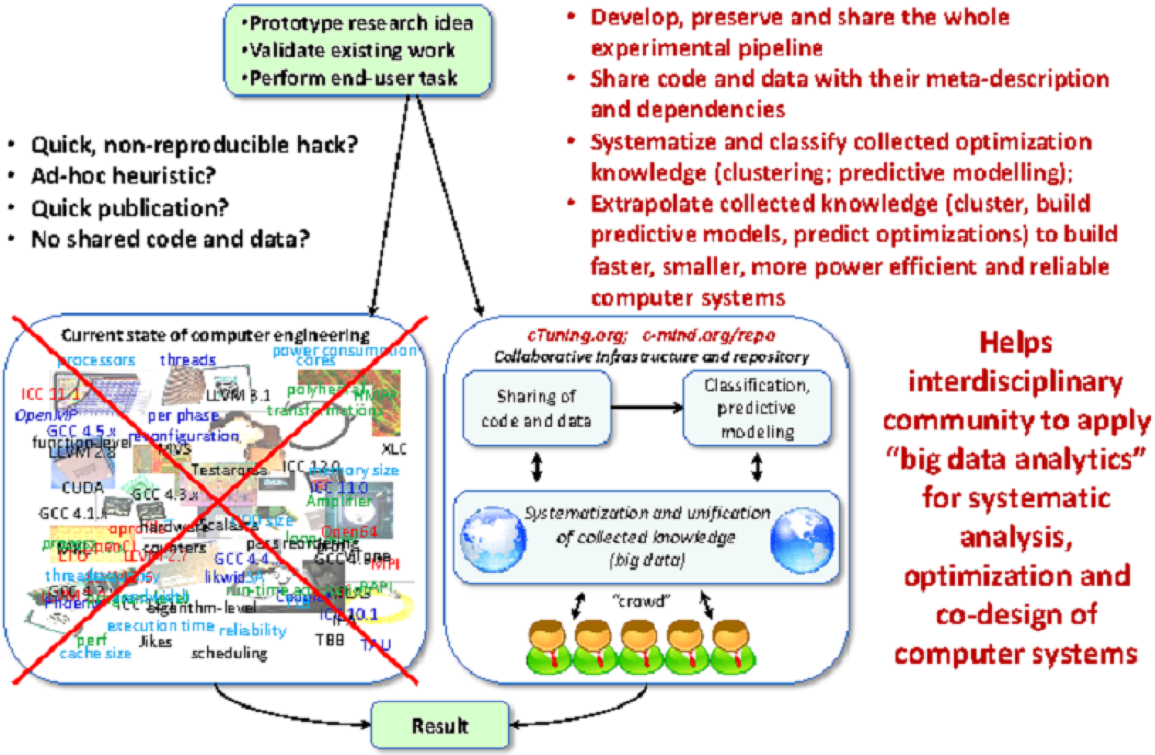
3.1.3.3. Research tracks for the 4 next years

In this project, we will be evaluating the first, to our knowledge, alternative, orthogonal, interdisciplinary, community-based and big-data driven approach to address above problems. We are developing a knowledge management system for computer engineering (possibly based on GPL-licensed cTuning and BSD-licensed Collective Mind) to preserve and share through the Internet the whole experimental (optimization) setups with all related artifacts and exposed meta-description in a unified way including behavior characteristics (execution time, code size, compilation time, power consumption, reliability, costs), semantic and dynamic features, design and optimization choices, and a system state together with all software and hardware dependencies besides just performance data. Such approach allows community to consider analysis, design and optimization of computer systems as a unified, formalized and big data problem while taking advantage of mature R&D methodologies from physics, biology and AI.

During this project, we will gradually structure, systematize, describe and share all research material in computer engineering including tools, benchmarks, data sets, search strategies and machine learning models. Researchers can later take advantage of shared components to collaboratively prototype, evaluate and improve various auto-tuning techniques while reusing all shared artifacts just like LEGO™pieces, and applying machine learning and data mining techniques to find meaningful relations between all shared material. It can also help crowdsource long tuning and learning process including classification and model building among many participants.

At the same time, any unexpected program behavior or model mispredictions can now be exposed to the community through unified web-services for collaborative analysis, explanation and solving. This, in turn, enables reproducibility of experimental results naturally and as a side effect rather than being enforced - interdisciplinary community needs to gradually find and add missing software and hardware dependencies to the Collective Mind (fixing processor frequency, pinning code to specific cores to avoid contentions) or improve analysis and predictive models (statistical normality tests for multiple experiments) whenever abnormal behavior is detected.

We hope that our approach will eventually help the community collaboratively evaluate and derive the most effective optimization strategies. It should also eventually help the community collaboratively learn complex behavior of all existing computer systems using top-down methodology originating from physics. At the same time, continuously collected and systematized knowledge ("big data") should allow community make quick and scientifically motivated advice about how to design and optimize the future heterogeneous HPC systems (particularly on our way towards extreme scale computing) as conceptually shown in Figure 3 .



- Prototype research idea
- Validate existing work
- Perform end-user task

- Quick, non-reproducible hack?
- Ad-hoc heuristic?
- Quick publication?
- No shared code and data?

- Develop, preserve and share the whole experimental pipeline
- Share code and data with their meta-description and dependencies
- Systematize and classify collected optimization knowledge (clustering; predictive modelling);
- Extrapolate collected knowledge (cluster, build predictive models, predict optimizations) to build faster, smaller, more power efficient and reliable computer systems

Current state of computer engineering

cTuning.org; c-mind.org/repo

Collaborative infrastructure and repository

Sharing of code and data

Classification, predictive modeling

Systematization and unification of collected knowledge (big data)

"crowd"

Helps interdisciplinary community to apply "big data analytics" for systematic analysis, optimization and co-design of computer systems

Result

Figure 3. *Considering program optimization and run-time adaptation as a "big data problem"*

Similar systematization, formalization and big data analytics already revolutionized biology, machine learning, robotics, AI, and other important scientific fields in the past decade. Our approach also started revolutionizing computer engineering making it more a science rather than non-systematic hacking. It helps us effectively deal with the rising complexity of computer systems while focusing on improving classification and predictive models of computer systems' behavior, and collaboratively find missing features (possibly using new deep learning algorithms and even unsupervised learning [92], [112]) to improve optimization predictions, rather than constantly reinventing techniques for each new program, architecture and environment.

Our approach is strongly supported by a recent Vinton G. Cerf's vision for computer engineering [59] as well as our existing technology, repository of knowledge and experience, and a growing community [77], [78], [79]. Even more importantly, our approach already helped to promote reproducible research and initiate a new publication model in computer engineering supported by ACM SIGPLAN where all experimental results and related research artifacts with their meta-description and dependencies are continuously shared along with publications to be validated and improved by the community [76].

3.2. High-level HPC libraries and applications

In this research topic, we focus on developing optimized algorithms and software for high-performance scientific computing and image processing.

3.2.1. Taking advantage of heterogeneous parallel architectures

3.2.1.1. Activity description

In recent years and as observed in the latest trends from the Top 500 list [0], heterogeneous computing combining manycore systems with accelerators such as Graphics Processing Units (GPU) or Intel Xeon Phi coprocessors has become a *de facto* standard in high performance computing. At the same time, data movements between memory hierarchies and/or between processors have become a major bottleneck for most numerical algorithms. The main goal of this topic is to investigate new approaches to develop linear algebra algorithms and software for heterogeneous architectures [42], [149], with also the objective of contributing to public domain numerical linear algebra libraries (e.g., MAGMA [0]).
Our activity in the field consists of designing algorithms that minimize the cost of communication and optimize data locality in numerical linear algebra solvers. When combining different architectures, these algorithms should be properly "hybridized". This means that the workload should be balanced throughout the execution, and the work scheduling/mapping should ensure matching of architectural features to algorithmic requirements.

In our effort to minimize communication, an example concerns the solution of general linear systems (via LU factorization) where the main objective is to reduce the communication overhead due to pivoting. We developed several algorithms to achieve this objective for hybrid CPU/GPU platforms. In one of them the panel factorization is performed using a communication-avoiding pivoting heuristic [83] while the update of the trailing submatrix is performed by the GPU [37]. In another algorithm, we use a random preconditioning (see also Section 3.2.2) of the original matrix to avoid pivoting [40]. Performance comparisons and tests on accuracy showed that these solvers are effective on current hybrid multicore-GPU parallel machines. These hybrid solvers will be integrated in a next release of the MAGMA library.

Another issue is related to the impact of non-uniform memory accesses (NUMA) on the solution of HPC applications. For dense linear systems, we illustrated how an appropriate placement of the threads and memory on a NUMA architecture can improve the performance of the panel factorization and consequently accelerate the global LU factorization [133], when compared to the hybrid multicore/GPU LU algorithm as it is implemented in the public domain library MAGMA.

[0] http://www.top500.org/
[0] Matrix Algebra on GPU and Multicore Architectures, http://icl.cs.utk.edu/magma/

3.2.1.2. Research tracks for the 4 next years

3.2.1.2.1. Towards automatic generation of dense linear solvers:

In an ongoing research, we investigate a generic description of the linear system to be solved in order to exploit numerical and structural properties of matrices to get fast and accurate solutions with respect to a specific type of problem. Information about targeted architectures and resources available will be also taken into account so that the most appropriate routines are used or generated. An application of this generative approach is the possibility of prototyping new algorithms or new implementations of existing algorithms for various hardware.

A track for generating efficient code is to develop new functionalities in the C++ library NT^2 [75] which is developed in the Postale team. This approach will enable us to generate optimized code that support current processor facilities (OpenMP and TBB support for multicores, SIMD extensions...) and accelerators (GPU, Intel Xeon Phi) starting from an API (Application Programming Interface) similar to Matlab. By analyzing the properties of the linear algebra domain that can be extracted from numerical libraries and combining them with architectural features, we have started to apply the generic approach mentioned in Section 3.1.2 to solve dense linear systems on various architectures including CPU and GPU. As an application, we plan to develop a new software that can run either on CPU or GPU to solve least squares problems based on semi-normal equations in mixed precision [36] since, to our knowledge, such a solver cannot be found in current public domain libraries (Sca)LAPACK [29], [54], PLASMA [150] and MAGMA [38]. This solver aims at attaining a performance that corresponds to what state-of-the-art codes achieve using mixed precision algorithms.

3.2.1.2.2. Communication avoiding algorithms for heterogeneous platforms:

In previous work, we focused on the LU decomposition with respect to two directions that are numerical stability and communication issue. This research work has lead to the development of a new algorithm for the LU decomposition, referred to as LU_PRRP: LU with panel rank revealing pivoting [98]. This algorithm uses a new pivoting strategy based on strong rank revealing QR factorization [84]. We also design a communication avoiding version of LU_PRRP, referred to as CALU_PRRP, which aims at overcoming the communication bottleneck during the panel factorization if we consider a parallel version of LU_PRRP. Thus CALU_PRRP is asymptotically optimal in terms of both bandwidth and latency. Moreover, it is more stable than the communication avoiding LU factorization based on Gaussian elimination with partial pivoting in terms of growth factor upper bound [64].

Due to the huge number and the heterogeneity of computing units in future exascale platforms, it is crucial for numerical algorithms to exhibit more parallelism and pipelining. It is thus important to study the critical paths of these algorithms, the task decomposition and the task granularity as well as the scheduling techniques in order to take advantage of the potential of the available platforms. Our goal here is to adapt our new algorithm CALU_PRRP to be scalable and efficient on heterogeneous platforms making use of the available accelerators and coprocessors similarly to what was achieved in [37].

3.2.1.2.3. Application to numerical fluid mechanics:

In an ongoing PhD thesis [153], [154], we apply hybrid programming techniques to develop a solver for the incompressible Navier-Stokes equations with constant coefficients, discretized by the finite difference method. In this application, we focus on solving large sparse linear systems coming from the discretization of Helmholtz and Poisson equations using direct methods that represent the major part of the computational time for solving the Navier-Stokes equations which describe a large class of fluid flows. In the future, our effort in the field will concern how to apply hybrid programming techniques to solvers based on iterative methods. A major task will consist of developing efficient kernels and choosing appropriate preconditioners. An important aspect is also the use of advanced scheduling techniques to minimize the number of synchronizations during the execution. The algorithms developed during this research activity will be validated on physical data provided by the physicists either form the academic world (e.g., LIMSI/University Paris-Sud [0] or industrial partners (e.g., EDF, ONERA). This research is currently performed in the framework of the CALIFHA project [0] and will be continued in an industrial contract with EDF R&D (starting October 2014).

[0] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, http://www.limsi.fr/

[0] CALculations of Incompressible Fluids on Heterogeneous, funded by Région Île-de-France and Digitéo (http://www.digiteo.fr)

3.2.2. Randomized algorithms in HPC applications

Activity description

Randomized algorithms are becoming very attractive in high-performance computing applications since they are able to outperform deterministic methods while still providing accurate results. Recent advances in the field include for instance random sampling algorithms [33], low-rank matrix approximation [116], or general matrix decompositions [87].

Our research in this domain consists of developing fast algorithms for linear algebra solvers which are at the heart of many HPC physical applications. In recent works, we designed randomized algorithms [40], [52] based on random butterfly transformations (RBT) [121] that can be applied to accelerate the solution of general or symmetric indefinite (dense) linear systems for multicore [35] or distributed architectures [34]. These randomized solvers have the advantage of reducing the amount of communication in dense factorizations by removing completely the pivoting phase which inhibits performance in Gaussian Elimination.

We also studied methods and software to assess the numerical quality of the solution computed in HPC applications. The objective is to compute quantities that provide us with information about the numerical quality of the computed solution in an acceptable time, at least significantly cheaper than the cost for the solution itself (typically a statistical estimation should require $\mathcal{O}(n^2)$ flops while the solution of a linear system involves at least $\mathcal{O}(n^3)$ flops, where n is the problem size). In particular, we recently applied in [44] statistical techniques based on the small sample theory [97] to estimate the condition number of linear system/linear least squares solvers [31], [39], [43]. This approach reduces significantly the number of arithmetic operations in estimating condition numbers. Whether designing fast solvers or error analysis tools, our ultimate goal is to integrate the resulting software into HPC libraries so that these routines will be available for physicists. The targeted architectures are multicore systems possibly accelerated with GPUs or Intel Xeon Phi coprocessors.

This research activity benefits from the Inria associate-team program, through the **associate-team R-LAS**[0], created in 2014 between Inria Saclay/Postale team and University of Tennessee (Innovative Computing Laboratory) in the area of randomized algorithms and software for numerical linear algebra. This project is funded from 2014 to 2016 and is lead jointly by Marc Baboulin (Inria/University Paris-Sud) and Jack Dongarra (University of Tennessee).

Research tracks for the 4 next years

3.2.2.1. Extension of random butterfly transformations to sparse matrices:

We recently illustrated how randomization via RBT can accelerate the solution of dense linear systems on multicore architectures possibly accelerated by GPUs. We recently started to extend this method to sparse linear systems arising from the discretization of partial differential equations in many physical applications. However, a major difficulty comes from the possible fill-in introduced by RBT. One of our first task consists of performing experiments on a collection of sparse matrices to evaluate the fill-in depending on the number of recursions in the algorithm. In a recent work [45], we investigated the possibility of using another form of RBT (one-side RBT instead of two-sided) in order to minimize the fill-in and we obtain promising preliminary results (Figure 4 shows that the fill-in is significantly reduced when using one-side RBT).

Another track of research is related to iterative methods for solving large sparse linear systems, and more particularly preconditioned Krylov subspace methods implemented in the solver ARMS (Algebraic Recursive Multilevel Solver (pARMS for its parallel distributed version). In this solver, our goal is to find the last level of preconditioning and then replace the original ILU factorization by our RBT preprocessing. A PhD thesis (supervised by Marc Baboulin) started in October 2014 on using randomization techniques like RBT for sparse linear systems.

[0] Randomized Linear Algebra Software, https://www.lri.fr/~baboulin/presentation_r-las.html/

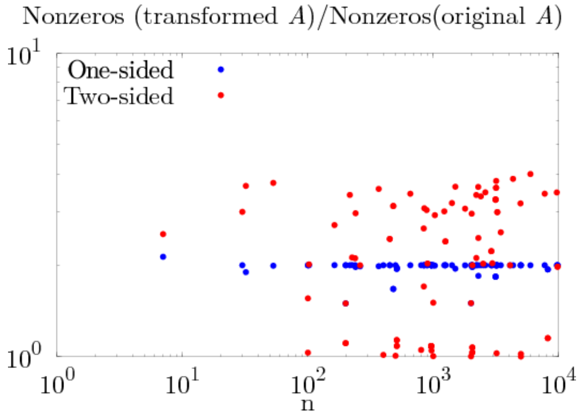


Figure 4. Evaluation of fill-in for one-sided RBT (90 matrices sorted by size).

3.2.2.2. Randomized algorithms on large clusters of multicore:

A major challenge for the randomized algorithms that we develop is to be able to solve very large problems arising in real-world physical simulations. As a matter of fact, large-scale linear algebra solvers from standard parallel distributed libraries like ScaLAPACK often suffer from expensive inter-node communication costs. An important requirement is to be able to schedule these algorithms dynamically on highly distributed and heterogeneous parallel systems [96]. In particular we point out that even though randomizing linear systems removes the communication due to pivoting, applying recursive butterflies also requires communication, especially if we use multiple nodes to perform the randomization. Our objective is to minimize this communication in the tiled agorithms and to use a runtime that enforces a strict data locality scheduling strategy [34]. A state of the art of possible runtime systems and how they can be combined with our randomized solvers will be established. Regarding the application of such solver, a collaboration with Pr Tetsuya Sakurai (University of Tsukuba, Japan) and Pr Jose Roman (Universitat Politècnica de València, Spain) will start in December 2014 to apply RBT to large linear systems encountered in contour integral eigensolver (CISS) [94]. Optimal tuning of the code will be obtained using holistic approach developed in the Postale team [79].

3.2.2.3. Extension of statistical estimation techniques to eigenvalue and singular value problems:

The extension of statistical condition estimation techniques can be carried out for eigenvalue/singular value calculations associated with nonsymmetric and symmetric matrices arising in, for example, optimization problems. In all cases, numerical sensitivity of the model parameters is of utmost concern and will guide the choice of estimation techniques. The important class of componentwise relative perturbations can be easily handled for a general matrix [97]. A significant outcome of the research will be the creation of high-quality open-source implementations of the algorithms developed in the project, similarly to the equivalent work for least squares problems [41]. To maximize its dissemination and impact, the software will be designed to be extensible, portable, and customizable.

3.2.2.4. Random orthogonal matrices:

Random orthogonal matrices have a wide variety of applications. They are used in the generation of various kinds of random matrices and random matrix polynomials [53], [63], [65], [91]. They are also used in some finance and statistics applications. For example the random orthogonal matrix (ROM) simulation [113] method uses random orthogonal matrices to generate multivariate random samples with the same mean and covariance as an observed sample.

The natural distribution over the space of orthogonal matrices is the Haar distribution. One way to generate a random orthogonal matrix from the Haar distribution is to generate a random matrix A with elements from the standard normal distribution and compute its QR factorization $A = QR$, where R is chosen to have nonnegative diagonal elements; the orthogonal factor Q is then the required matrix [90].

Stewart [140] developed a more efficient algorithm that directly generates an $n \times n$ orthogonal matrix from the Haar distribution as a product of Householder transformations built from Householder vectors of dimensions $1, 2, \cdots, n-1$ chosen from the standard normal distribution. Our objective is to design an algorithm that significantly reduces the computational cost of Stewart's algorithm by relaxing the property that Q is exactly Haar distributed. We also aim at extending the use of random orthogonal matrices to other randomized algorithms.

3.2.3. Embedded high-performance systems & computer vision

Scientific context

High-performance embedded systems & computer vision address the design of efficient algorithms for parallel architectures that deal with image processing and computer vision. Such systems must enforce realtime execution constraint (typically 25 frames per second) and power consumption constraint. If no COTS (*Component On The Shelf*) architecture (e.g., SIMD multicore processor, GPU, Intel Xeon Phi, DSP) satisfy the constraints, then we have to develop a specialized one.

A more and more important aspect when designing an embedded system is the tradeoff between speed (and power consumption) and numerical accuracy (and stability). Such a tradeoff leads to 16-bit computation (and storage) and to the design of less accurate algorithms. For example, the final accuracy for stabilizing an image is $10-1$ pixel, which is far from the maximum accuracy of (10^{-7}) available using the 32-bit IEEE format.

3.2.3.1. Activity description and recent achievements

Concerning image processing, our efforts concern the redesign of data-dependent algorithms for parallel architectures. A representative example of such an algorithm is the connected-component labeling (CCL) algorithm [132] which is used in industrial or medical imaging and classical computer vision like optical character recognition. As far as we know our algorithm (*Light Speed Labeling*) [57], [58] still outperforms other existing CCL algorithms [82], [89], [145] (the first versions of our algorithm appeared in 2009 [105], [106]).

Concerning computer vision (smart camera, autonomous robot, aerial drone), we developed in collaboration with LIMSI [0] two applications that run in realtime on embedded parallel systems [107], [131] with some accuracy tradeoffs. The first one is based on mean shift tracking [80], [81] and the second one relies on covariance matching and tracking [128], [129], [130].

These applications are used in video-surveillance: they perform motion detection [104], motion analysis [146], [147], motion estimation and multi-target tracking. Depending on the image nature and size, some algorithmic transforms (integral image, cumulative differential sum) can be applied and combined with hybrid arithmetic (16-bit / 32-bit / 64-bit). Finally, to increase the algorithm robustness color, space optimization is also used [108].

Usually one tries to convert 64-bit computations into 32-bit. But sometimes 16-bit floating point arithmetic is sufficient. As 16-bit numbers are now normalized by IEEE (754-2008) and are available in COTS processors like GPU and GPP (AVX2 for storage in memory and conversion into 32-bit numbers), we can run such kind of code on COTS processors or we can design specialized architectures like FPGA (*Field-Programmable gate array*) and ASIC (*Application-specific integrated circuit*) to be more efficient. This approach is complementary to that of [117] which converts 32-bit floating point signal processing operators into fixed-point ones.

[0] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

By extension to computer vision, we also address *interactive sensing HPC applications*. One CEA thesis funded by CEA and co-supervised by Lionel Lacassagne addresses the parallelization of Non Destructive Testing applications on COTS processors (super-charged workstation with GPUs and Intel Xeon Phi manycore processor). This PhD thesis deals with irregular computations with sparse-addressing and load-balacing problems. It also deals with floating point accuracy, by finding roots of polynomials using Newton and Laguerre algorithms. Depending on the configuration, 64-bit is required, but sometime 32-bit computations are sufficient with respect to the physics. As the second application focuses on interactive sensing, one has to add a second level of tradeoff for physical sampling accuracy and the sensor displacement [109], [110], [111], [126], [127].
In order to achieve realtime execution on the targeted architectures, we develop *High Level Transforms* (HLT) that are algorithmic transforms for memory layout and function re-organization. We show on a representative algorithm [88] in the image processing area that a fully parallelized code (SIMD+OpenMP) can be accelerated by a factor ×80 on a multicore processor [101]. A CIFRE thesis (defended in 2014) funded by ST Microelectronics and supervised by Lionel Lacassagne has led to the design of very efficient implementations into an ASIC thanks to HLT. We show that the power consumption can be reduced by a factor 10 [155], [156].
All these applications have led to the development of software libraries for image processing that are currently under registration at APP (Agence de Protection des Programmes): `myNRC 2.0`[0] and `covTrack`[0].

3.2.3.2. Future: system, image & arithmetic

Concerning image processing we are designing new versions of CCL algorithms. One version is for parallel architectures where graph merging and efficient transitive closure is a major issue for load balancing. For embedded systems, *time prediction* is as important as execution time, so a specialized version targets embedded processors like ARM processors and Texas Instrument VLIW DSP C6x.

We also plan to design algorithms that should be less data-sensitive (the execution time depends on the nature of the image: a structured image can be processed quickly whereas an unstructured image will require more time). These algorithms will be used in even more data-dependent algorithms like *hysteresis thresholding* for image binarization, *split-and-merge* [30], [100] for realtime image segmentation using the Horowitz-Pavlidis quad-tree decomposition [93]. Such an algorithm could be useful for accelerating image decomposition like *Fast Level Set Transform* algorithm [118].

Concerning Computer vision we will study 16-bit floating point arithmetic for image processing applications and linear algebra operators. Concerning image processing, we will focus on iterative algorithms like optical flow computation (for motion estimation and image stabilization). We will compare the efficiency (accuracy and speed) of 16-bit floating point [72], [103], [102], [124] with fixed-point arithmetic. Concerning linear algebra, we will study efficient implementation for very small matrix inversion (from 6×6 up to 16×16) for our covariance-tracking algorithm.

According to Nvidia (see Figure 5), the computation rate (Gflop/s) for ZGEMM (complex matrix-matrix multiplication with 64-bit precision – for small value of N – is linearly proportional to N. That means that, for a 6×6 matrix, we achieve around 6 Gflop/s on a Tesla M2090 (400 Gflop/s peak power). This represents 1.5 % of the peak power. For that reason, designing efficient parallel codes for embedded systems [60], [67], [68] is different and may be more complex than designing codes for classical HPC systems. Our `covTrack` software requires many hundreds of 6×6 matrix-matrix multiplications every frame.
Last point is to develop tools that help to automatically distribute or parallelize a code on an architecture code parallelization/distribution dealing with scientific computing [69], MPI [73] or image applications on the Cell processor [61], [74], [123], [134], [135], [136], [148].

[0] smart memory allocator and management for 2D and 3D image processing

[0] agile realtime multi-target tracking algorithm, co-developped with Michèle Gouiffès at LIMSI

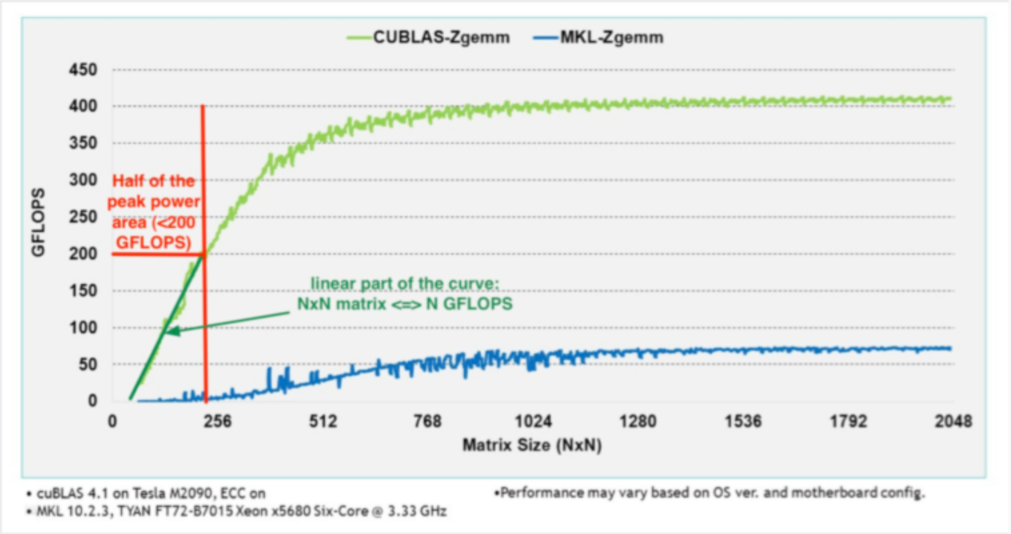


Figure 5. Nvidia cuBLAS performance versus Intel MKL: both have poor performance for small N

SPECFUN Project-Team

3. Research Program

3.1. Studying special functions by computer algebra

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

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

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

It aims at four kinds of algorithmic goals:

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

This interacts with three domains of research:

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

This view is made explicit in the present section.

3.1.1. Equations as a data structure

Numerous special functions satisfy linear differential and/or recurrence equations. Under a mild technical condition, the existence of such equations induces a finiteness property that makes the main properties of the functions decidable. We thus speak of *D-finite functions*. For example, 60 % of the chapters in the handbook [15] describe D-finite functions. In addition, the class is closed under a rich set of algebraic operations. This makes linear functional equations just the right data structure to encode and manipulate special functions. The power of this representation was observed in the early 1990s [69], leading to the design of many algorithms in computer algebra. Both on the theoretical and algorithmic sides, the study of D-finite functions shares much with neighbouring mathematical domains: differential algebra, D-module theory, differential Galois theory, as well as their counterparts for recurrence equations.

3.1.2. Algorithms combining functions

Differential/recurrence equations that define special functions can be recombined [69] to define: additions and products of special functions; compositions of special functions; integrals and sums involving special functions. Zeilberger’s fast algorithm for obtaining recurrences satisfied by parametrised binomial sums was developed in the early 1990s already [70]. It is the basis of all modern definite summation and integration algorithms. The theory was made fully rigorous and algorithmic in later works, mostly by a group in RISC (Linz, Austria) and by members of the team [58], [66], [34], [32], [33], [52]. The past ÉPI Algorithms contributed several implementations (*gfun* [61], *Mgfun* [34]).

3.1.3. Solving functional equations

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

3.1.4. Multi-precision numerical evaluation

A major conceptual and algorithmic difference exists for numerical calculations between data structures that fit on a machine word and data structures of arbitrary length, that is, *multi-precision* arithmetic. When multi-precision floating-point numbers became available, early works on the evaluation of special functions were just promising that "most" digits in the output were correct, and performed by heuristically increasing precision during intermediate calculations, without intended rigour. The original theory has evolved in a twofold way since the 1990s: by making computable all constants hidden in asymptotic approximations, it became possible to guarantee a *prescribed* absolute precision; by employing state-of-the-art algorithms on polynomials, matrices, etc, it became possible to have evaluation algorithms in a time complexity that is linear in the output size, with a constant that is not more than a few units. On the implementation side, several original works exist, one of which (*NumGfun* [57]) is used in our DDMF.

3.1.5. Guessing heuristics

"Differential approximation", or "Guessing", is an operation to get an ODE likely to be satisfied by a given approximate series expansion of an unknown function. This has been used at least since the 1970s and is a key stone in spectacular applications in experimental mathematics [30]. All this is based on subtle algorithms for Hermite–Padé approximants [19]. Moreover, guessing can at times be complemented by proven quantitative results that turn the heuristics into an algorithm [27]. This is a promising algorithmic approach that deserves more attention than it has received so far.

3.1.6. Complexity-driven design of algorithms

The main concern of computer algebra has long been to prove the feasibility of a given problem, that is, to show the existence of an algorithmic solution for it. However, with the advent of faster and faster computers, complexity results have ceased to be of theoretical interest only. Nowadays, a large track of works in computer algebra is interested in developing fast algorithms, with time complexity as close as possible to linear in their output size. After most of the more pervasive objects like integers, polynomials, and matrices have been endowed with fast algorithms for the main operations on them [39], the community, including ourselves, started to turn its attention to differential and recurrence objects in the 2000s. The subject is still not as developed as in the commutative case, and a major challenge remains to understand the combinatorics behind summation and integration. On the methodological side, several paradigms occur repeatedly in fast algorithms: "divide and conquer" to balance calculations, "evaluation and interpolation" to avoid intermediate swell of data, etc. [24].

3.2. Trusted computer-algebra calculations

3.2.1. Encyclopedias

Handbooks collecting mathematical properties aim at serving as reference, therefore trusted, documents. The decision of several authors or maintainers of such knowledge bases to move from paper books [15], [17], [62] to websites and wikis [0] allows for a more collaborative effort in proof reading. Another step toward further confidence is to manage to generate the content of an encyclopedia by computer-algebra programs, as is the case with the Wolfram Functions Site [0] or DDMF [0]. Yet, due to the lingering doubts about computer-algebra systems, some encyclopedias propose both cross-checking by different systems and handwritten companion paper proofs of their content [0]. As of today, there is no encyclopedia certified with formal proofs.

[0] for instance http://dlmf.nist.gov/ for special functions or http://oeis.org/ for integer sequences

[0] http://functions.wolfram.com/

[0] http://ddmf.msr-inria.inria.fr/1.9.1/ddmf

3.2.2. Computer algebra and symbolic logic

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

3.2.3. Certifying systems for computer algebra

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

3.2.4. Semantics for computer algebra

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

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

A number of authors have investigated ways to organize the communication of a chosen computer-algebra system with a chosen proof assistant in order to certify specific components of the computer-algebra systems, experimenting various combinations of systems and various formats for mathematical exchanges. Another line of research consists in the implementation and certification of computer-algebra algorithms inside the logic [65], [44], [54] or as a proof-automation strategy. Normalization algorithms are of special interest when they allow to check results possibly obtained by an external computer-algebra oracle [37]. A discussion about the systematic separation of the search for a solution and the checking of the solution is already clearly outlined in [50].

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

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

3.3. Machine-checked proofs of formalized mathematics

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

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

3.3.1. Logical foundations and proof assistants

While many logical foundations for mathematics have been proposed, studied, and implemented, type theory is the one that has been more successfully employed to formalize mathematics, to the notable exception of the Mizar system [55], which is based on set theory. In particular, the calculus of construction (CoC) [35] and its extension with inductive types (CIC) [36], have been studied for more than 20 years and been implemented by several independent tools (like Lego, Matita, and Agda). Its reference implementation, Coq [63], has been used for several large-scale formalizations projects (formal certification of a compiler back-end; four-color theorem). Improving the type theory underlying the Coq system remains an active area of research. Other systems based on different type theories do exist and, whilst being more oriented toward software verification, have been also used to verify results of mainstream mathematics (prime-number theorem; Kepler conjecture).

3.3.2. Computations in formal proofs

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

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

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

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

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

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

3.3.4. Relevant contributions from the Mathematical Component libraries

Representing abstract algebra in a proof assistant has been studied for long. The libraries developed by the MathComp project for the proof of the Odd Order Theorem provide a rather comprehensive hierarchy of structures; however, they originally feature a large number of instances of structures that they need to organize. On the methodological side, this hierarchy is an incarnation of an original work [38] based on various mechanisms, primarily type inference, typically employed in the area of programming languages. A large amount of information that is implicit in handwritten proofs, and that must become explicit at formalization time, can be systematically recovered following this methodology.

Small-scale reflection [41] is another methodology promoted by the MathComp project. Its ultimate goal is to ease formal proofs by systematically dealing with as many bureaucratic steps as possible, by automated computation. For instance, as opposed to the style advocated by Coq's standard library, decidable predicates are systematically represented using computable boolean functions: comparison on integers is expressed as program, and to state that $a \leq b$ one compares the output of this program run on a and b with $true$. In many cases, for example when a and b are values, one can prove or disprove the inequality by pure computation.

The MathComp library was consistently designed after uniform principles of software engineering. These principles range from simple ones, like naming conventions, to more advanced ones, like generic programming, resulting in a robust and reusable collection of formal mathematical components. This large body of formalized mathematics covers a broad panel of algebraic theories, including of course advanced topics of finite group theory, but also linear algebra, commutative algebra, Galois theory, and representation theory. We refer the interested reader to the online documentation of these libraries [64], which represent about 150,000 lines of code and include roughly 4,000 definitions and 13,000 theorems.

Topics not addressed by these libraries and that might be relevant to the present project include real analysis and differential equations. The most advanced work of formalization on these domains is available in the HOL-Light system [46], [47], [48], although some existing developments of interest [22], [56] are also available for Coq. Another aspect of the MathComp libraries that needs improvement, owing to the size of the data we manipulate, is the connection with efficient data structures and implementations, which only starts to be explored.

3.3.5. User interaction with the proof assistant

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

The development of the large, consistent body of MathComp libraries has prompted the need to design an alternative and coherent language extension for the Coq proof assistant [43], [42], enforcing the robustness of proof scripts to the numerous changes induced by code refactoring and enhancing the support for the methodology of small-scale reflection.

The development of large libraries is quite a novelty for the Coq system. In particular any long-term development process requires the iteration of many refactoring steps and very little support is provided by most proof assistants, with the notable exception of Mizar [60]. For the Coq system, this is an active area of research.

TOCCATA Project-Team

3. Research Program

3.1. Introduction

In the former ProVal project, we have been working on the design of methods and tools for deductive verification of programs. One of our original skills was the ability to conduct proofs by using automatic provers and proof assistants at the same time, depending on the difficulty of the program, and specifically the difficulty of each particular verification condition. We thus believe that we are in a good position to propose a bridge between the two families of approaches of deductive verification presented above. Establishing this bridge is one of the goals of the Toccata project: we want to provide methods and tools for deductive program verification that can offer both a high amount of proof automation and a high guarantee of validity. Toward this objective, a new axis of research was proposed: the development of *certified* analysis tools that are themselves formally proved correct.

The reader should be aware that the word “certified” in this scientific programme means “verified by a formal specification and a formal proof that the program meets this specification”. This differs from the standard meaning of “certified” in an industrial context where it means a conformance to some rigorous process and/or norm. We believe this is the right term to use, as it was used for the *Certified Compiler* project [96], the new conference series *Certified Programs and Proofs*, and more generally the important topics of *proof certificates*.

In industrial applications, numerical calculations are very common (e.g. control software in transportation). Typically they involve floating-point numbers. Some of the members of Toccata have an internationally recognized expertise on deductive program verification involving floating-point computations. Our past work includes a new approach for proving behavioral properties of numerical C programs using Frama-C/Jessie [37], various examples of applications of that approach [56], the use of the Gappa solver for proving numerical algorithms [113], an approach to take architectures and compilers into account when dealing with floating-point programs [57], [107]. We also contributed to the Handbook of Floating-Point Arithmetic [105]. A representative case study is the analysis and the proof of both the method error and the rounding error of a numerical analysis program solving the one-dimension acoustic wave equation [5] [50]. Our experience led us to a conclusion that verification of numerical programs can benefit a lot from combining automatic and interactive theorem proving [52], [56]. Certification of numerical programs is the other main axis of Toccata.

Our scientific programme in structured into four objectives:

1. deductive program verification;
2. automated reasoning;
3. formalization and certification of languages, tools and systems;
4. proof of numerical programs.

We detail these objectives below.

3.2. Deductive Program Verification

Permanent researchers: A. Charguéraud, S. Conchon, J.-C. Filliâtre, C. Marché, G. Melquiond, A. Paskevich

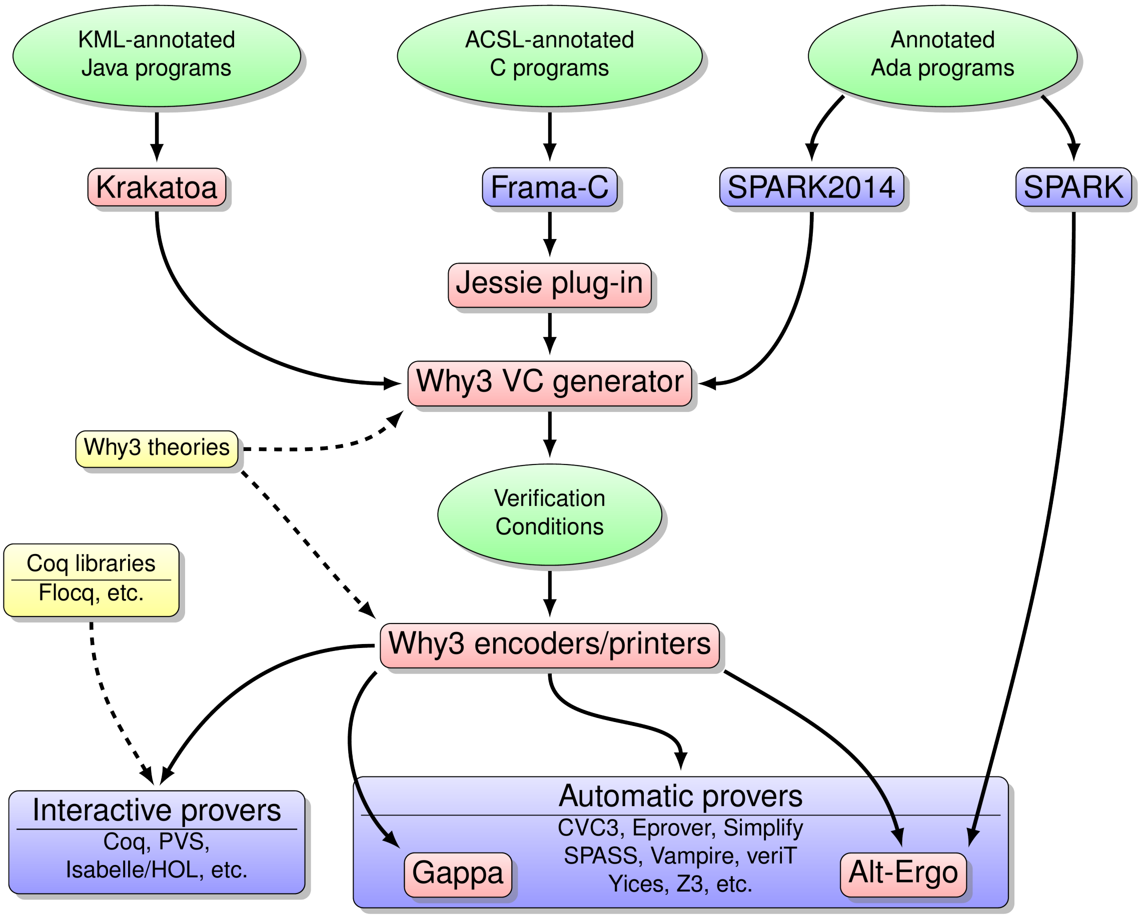


Figure 1. The Why3 ecosystem

3.2.1. The Why3 Ecosystem

This ecosystem is central in our work; it is displayed on Figure 1 . The boxes in red background correspond to the tools we develop in the Toccata team.

- The initial design of Why3 was presented in 2012 [45], [85]. In the past years, the main improvements concern the specification language (such as support for higher-order logic functions [63]) and the support for provers. Several new interactive provers are now supported: PVS 6 (used at NASA), Isabelle2014 (planned to be used in the context of Ada program via Spark), and Mathematica. We also added support for new automated provers: CVC4, Metitarski, Metis, Beagle, Princess, and Yices2. More technical improvements are the design of a Coq tactic to call provers via Why3 from Coq, and the design of a proof session mechanism [44]. Why3 was presented during several invited talks [84], [83], [80], [81].
- J.-C. Filliâtre, L. Gondelman, and A. Paskevich have designed a general approach to the concept of ghost code [9] which is a subset of program code that serves the purposes of specification and verification: it can be erased from the program without affecting its result. This work forms the basis of the support for ghost code in Why3.
- At the level of the C front-end of Why3 (via Frama-C), we have proposed an approach to add a notion of refinement on C programs [112], and an approach to reason about pointer programs with a standard logic, via *separation predicates* [43]
- The Ada front-end of Why3 has mainly been developed during the past three years, leading to the release of SPARK2014 [91] (http://www.spark-2014.org/)
- In collaboration with J. Almeida, M. Barbosa, J. Pinto, and B. Vieira (University do Minho, Braga, Portugal), J.-C. Filliâtre has developed a method for certifying programs involving cryptographic methods. It uses Why as an intermediate language [36].
- With M. Pereira and S. Melo de Sousa (Universidade da Beira Interior, Covilhã, Portugal), J.-C. Filliâtre has developed an environment for proving ARM assembly code. It uses Why3 as an intermediate VC generator. It was presented at the Inforum conference [110] (best student paper).

3.2.2. Analysis of Complexity

- A. Charguéraud has recently extended his tool CFML [61] to support, in addition to the verification of the full functional correctness of a piece of code, the verification of the asymptotic complexity of the code [24]. Even though it had been previously established that, in theory, amortized analysis can be explained as the manipulation of *time credits*, and that time credits can be encoded as resources in Separation Logic, CFML is the first practical tool to support the formal verification of amortized analyses for arbitrarily complex pieces of code.

3.2.3. Concurrent Programming

- S. Conchon and A. Mebsout, in collaboration with F. Zaïdi (VALS team, LRI), A. Goel and S. Krstić (Strategic Cad Labs, INTEL) have proposed a new model-checking approach for verifying safety properties of array-based systems. This is a syntactically restricted class of parametrized transition systems with states represented as arrays indexed by an arbitrary number of processes. Cache coherence protocols and mutual exclusion algorithms are typical examples of such systems. It was first presented at CAV 2012 [8] and detailed further [72]. It was applied to the verification of programs with fences [68]. The core algorithm has been extended with a mechanism for inferring invariants. This new algorithm, called BRAB, is able to automatically infer invariants strong enough to prove industrial cache coherence protocols. BRAB computes over-approximations of backward reachable states that are checked to be unreachable in a finite instance of the system. These approximations (candidate invariants) are then model-checked together with the original safety properties. Completeness of the approach is ensured by a mechanism for backtracking on spurious traces introduced by too coarse approximations [69], [101].

- In the context of the ERC DeepSea project [0], A. Charguéraud and his co-authors have developed a load balancing algorithm that implements the work stealing scheme using private deques [34]. They have shown this algorithm to be implementable even without atomic operations on x86-TSO architectures, and established (on paper) the correctness of the algorithm using a novel proof technique [33]. They have also developed a *chunked sequence* data structure [35] that supports logarithmic concatenation and splitting, motivated by application to parallelism. In particular, A. Charguéraud and his co-authors have built, on top of the aforementioned work, fast and robust parallel graph traversal algorithms for parallel BFS and parallel DFS [20], [29].

3.2.4. Case Studies

- To provide an easy access to the case studies that we develop using Why3 and its front-ends, we have published a *gallery of verified programs* on our web page http://toccata.lri.fr/gallery/. Part of these examples are the solutions to the competitions VerifyThis 2011 [58], VerifyThis 2012 [12], and the competition VScomp 2011 [86].
- Other case studies that led to publications are the design of a library of data-structures based on AVLs [62], and the verification a two-lines C program (solving the N-queens puzzle) using Why3 [82].
- A. Charguéraud put the new *time credit* extension of CFML to practice to verify the correctness and asympotic complexity of a *chunked sequence* data structure [35], particularly challenging due to its use of Tarjan's data structural bootstrapping technique. Furthermore, A. Charguéraud and F. Pottier applied they same approach to formalize an efficient implementation of the classic Union Find data structure, which features the bound expressed in terms of the inverse Ackermann function [24].

For other case studies, see also sections of numerical programs and formalization of languages and tools.

3.2.5. Project-team Positioning

Several research groups in the world develop their own approaches, techniques, and tools for deductive verification. With respect to all these related approaches and tools, our originality is our will to use more sophisticated specification languages (with inductive definitions, higher-order features and such) and the ability to use a large set of various theorem provers, including the use of interactive theorem proving to deal with complex functional properties.

- The RiSE team [0] at Microsoft Research Redmond, USA, partly in collaboration with team "programming methodology" team [0] at ETH Zurich develop tools that are closely related to ours: Boogie and Dafny are direct competitors of Why3, VCC is a direct competitor of Frama-C/Jessie.
- The KeY project [0] (several teams, mainly at Karlsruhe and Darmstadt, Germany, and Göteborg, Sweden) develops the KeY tool for Java program verification [32], based on dynamic logic, and has several industrial users. They use a specific modal logic (dynamic logic) for modeling programs, whereas we use standard logic, so as to be able to use off-the-shelf automated provers.
- The "software engineering" group at Augsburg, Germany, develops the KIV system [0], which was created more than 20 years ago (1992) and is still well maintained and efficient. It provides a semi-interactive proof environment based on algebraic-style specifications, and is able to deal with several kinds of imperative style programs. They have a significant industrial impact.
- The VeriFast system [0] aims at verifying C programs specified in Separation Logic. It is developed at the Katholic University at Leuven, Belgium. We do not usually use separation logic (so as to use off-the-shelf provers) but alternative approaches (e.g. static memory separation analysis).

[0] Arthur Charguéraud is involved 40% of his time in the ERC DeepSea project, which is hosted at Inria Paris Rocquencourt (team Gallium).

[0] http://research.microsoft.com/en-us/groups/rise/default.aspx

[0] http://www.pm.inf.ethz.ch/

[0] http://www.key-project.org/

[0] http://www.isse.uni-augsburg.de/en/software/kiv/

[0] http://people.cs.kuleuven.be/~bart.jacobs/verifast/

- The Mobius Program Verification Environment [0] is a joint effort for the verification of Java source annotated with JML, combining static analysis and runtime checking. The tool ESC/Java2 [0] is a VC generator similar to Krakatoa, that builds on top of Boogie. It is developed by a community leaded by University of Copenhagen, Denmark. Again, our specificity with respect to them is the consideration of more complex specification languages and interactive theorem proving.
- The Lab for Automated Reasoning and Analysis [0] at EPFL, develop methods and tools for verification of Java (Jahob) and Scala (Leon) programs. They share with us the will and the ability to use several provers at the same time.
- The TLA environment [0], developed by Microsoft Research and the Inria team Veridis, aims at the verification of concurrent programs using mathematical specifications, model checking, and interactive or automated theorem proving.
- The F* project [0], developed by Microsoft Research and the Inria Prosecco team, aims at providing a rich environment for developing programs and proving them.

The KeY and KIV environments mentioned above are partly based on interactive theorem provers. There are other approaches on top of general-purpose proof assistants for proving programs that are not purely functional:

- The Ynot project [0] is a Coq library for writing imperative programs specified in separation logic. It was developed at Harvard University, until the end of the project in 2010. Ynot had similar goals as CFML, although Ynot requires programs to be written in monadic style inside Coq, whereas CFML applies directly on programs written in OCaml syntax, translating them into logical formulae.
- Front-ends to Isabelle were developed to deal with simple sequential imperative programs [111] or C programs [109]. The L4-verified project [92] is built on top of Isabelle.

3.3. Automated Reasoning

Permanent researchers: S. Conchon, É. Contejean, G. Melquiond, A. Paskevich

3.3.1. Generalities on Automated Reasoning

- J. C. Blanchette and A. Paskevich have designed an extension to the TPTP TFF (Typed First-order Form) format of theorem proving problems to support rank-1 polymorphic types (also known as ML-style parametric polymorphism) [2]. This extension, named TFF1, has been incorporated in the TPTP standard.
- S. Conchon defended his *habilitation à diriger des recherches* in December 2012. The memoir [65] provides a useful survey of the scientific work of the past 10 years, around the SMT solving techniques, that led to the tools Alt-Ergo and Cubicle as they are nowadays.
- É. Contejean [74] defended her *habilitation à diriger des recherches* in June 2014, where she examined the field of deduction from multiple points of view: theoretical and practical, automated, as well as interactive.

3.3.2. Quantifiers and Triggers

- C. Dross, J. Kanig, S. Conchon, and A. Paskevich have proposed a generic framework for adding a decision procedure for a theory or a combination of theories to an SMT prover. This mechanism is based on the notion of instantiation patterns, or *triggers*, which restrict instantiation of universal premises and can effectively prevent a combinatorial explosion. A user provides an axiomatization with triggers, along with a proof of completeness and termination in the proposed framework, and

[0] http://kindsoftware.com/products/opensource/Mobius/
[0] http://kindsoftware.com/products/opensource/ESCJava2/
[0] http://lara.epfl.ch/w/
[0] http://research.microsoft.com/en-us/um/people/lamport/tla/tla.html
[0] http://research.microsoft.com/en-us/projects/fstar/
[0] http://ynot.cs.harvard.edu/

obtains in return a sound, complete and terminating solver for his theory. A prototype implementation was realized on top of Alt-Ergo. As a case study, a feature-rich axiomatization of doubly-linked lists was proved complete and terminating [78]. C. Dross defended her PhD thesis in April 2014 [79]. The main results of the thesis are: (1) a formal semantics of the notion of *triggers* typically used to control quantifier instantiation in SMT solvers, (2) a general setting to show how a first-order axiomatization with triggers can be proved correct, complete, and terminating, and (3) an extended DPLL(T) algorithm to integrate a first-order axiomatization with triggers as a decision procedure for the theory it defines. Significant case studies were conducted on examples coming from SPARK programs, and on the benchmarks on B set theory constructed within the BWare project.

3.3.3. Reasoning Modulo Theories

- S. Conchon, É. Contejean and M. Iguernelala have presented a modular extension of ground AC-completion for deciding formulas in the combination of the theory of equality with user-defined AC symbols, uninterpreted symbols and an arbitrary signature-disjoint Shostak theory X [67]. This work extends the results presented in [66] by showing that a simple preprocessing step allows to get rid of a full AC-compatible reduction ordering, and to simply use a partial multiset extension of a *non-necessarily AC-compatible* ordering.
- S. Conchon, M. Iguernelala, and A. Mebsout have designed a collaborative framework for reasoning modulo simple properties of non-linear arithmetic [71]. This framework has been implemented in the Alt-Ergo SMT solver.
- S. Conchon, G. Melquiond and C. Roux have described a dedicated procedure for a theory of floating-point numbers which allows reasoning on approximation errors. This procedure is based on the approach of the Gappa tool: it performs saturation of consequences of the axioms, in order to refine bounds on expressions. In addition to the original approach, bounds are further refined by a constraint solver for linear arithmetic [73]. This procedure has been implemented in Alt-Ergo.
- In collaboration with A. Mahboubi (Inria project-team Typical), and G. Melquiond, the group involved in the development of Alt-Ergo have implemented and proved the correctness of a novel decision procedure for quantifier-free linear integer arithmetic [3]. This algorithm tries to bridge the gap between projection and branching/cutting methods: it interleaves an exhaustive search for a model with bounds inference. These bounds are computed provided an oracle capable of finding constant positive linear combinations of affine forms. An efficient oracle based on the Simplex procedure has been designed. This algorithm is proved sound, complete, and terminating and is implemented in Alt-Ergo.
- Most of the results above are detailed in M. Iguernelala's PhD thesis [89].

3.3.4. Applications

- We have been quite successful in the application of Alt-Ergo to industrial development: qualification by Airbus France, integration of Alt-Ergo into the Spark Pro toolset.
- In the context of the BWare project, aiming at using Why3 and Alt-Ergo for discharging proof obligations generated by Atelier B, we made progress into several directions. The method of translation of B proof obligations into Why3 goals was first presented at ABZ'2012 [104]. Then, new drivers have been designed for Why3, in order to use new back-end provers Zenon modulo and iProver modulo. A notion of rewrite rule was introduced into Why3, and a transformation for simplifying goals before sending them to back-end provers was designed. Intermediate results obtained so far in the project were presented both at the French conference AFADL [77] and at ABZ'2014 [76].

  On the side of Alt-Ergo, recent developments have been made to efficiently discharge proof obligations generated by Atelier B. This includes a new plugin architecture to facilitate experiments with different SAT engines, new heuristics to handle quantified formulas, and important modifications in its internal data structures to boost performances of core decision procedures. Benchmarks realized on more than 10,000 proof obligations generated from industrial B projects show significant improvements [70].

- Hybrid automatons interleave continuous behaviors (described by differential equations) with discrete transitions. D. Ishii and G. Melquiond have worked on an automated procedure for verifying safety properties (that is, global invariants) of such systems [90].

3.3.5. Project-team Positioning

Automated Theorem Proving is a large community, but several sub-groups can be identified:

- The SMT-LIB community gathers people interested in reasoning modulo theories. In this community, only a minority of participants are interested in supporting first-order quantifiers at the same time as theories. SMT solvers that support quantifiers are Z3 (Microsoft Research Redmond, USA), CVC3 and its successor CVC4 [0].
- The TPTP community gathers people interested in first-order theorem proving.
- Other Inria teams develop provers: veriT by team Veridis, and Psyche by team Parsifal.
- Other groups develop provers dedicated to very specific cases, such as Metitarski [0] at Cambridge, UK, which aims at proving formulas on real numbers, in particular involving special functions such as log or exp. The goal is somewhat similar to our CoqInterval library, *cf* objective 4.

It should be noticed that a large number of provers mentioned above are connected to Why3 as back-ends.

3.4. Formalization and Certification of Languages, Tools and Systems

Permanent researchers: S. Boldo, A. Charguéraud, É. Contejean, C. Marché, G. Melquiond, C. Paulin

3.4.1. Real Numbers, Real Analysis, Probabilities

- S. Boldo, C. Lelay, and G. Melquiond have worked on the Coquelicot library, designed to be a user-friendly Coq library about real analysis [54][14]. An easier way of writing formulas and theorem statements is achieved by relying on total functions in place of dependent types for limits, derivatives, integrals, power series, and so on. To help with the proof process, the library comes with a comprehensive set of theorems and some automation. We have exercised the library on several use cases: on an exam at university entry level [94], for the definitions and properties of Bessel functions [93], and for the solution of the one-dimensional wave equation [95]. We have also conducted a survey on the formalization of real arithmetic and real analysis in various proof systems [55].
- Watermarking techniques are used to help identify copies of publicly released information. They consist in applying a slight and secret modification to the data before its release, in a way that should remain recognizable even in (reasonably) modified copies of the data. Using the Coq ALEA library, which formalizes probability theory and probabilistic programs, D. Baelde together with P. Courtieu, D. Gross-Amblard from Rennes and C. Paulin have established new results about the robustness of watermarking schemes against arbitrary attackers [38]. The technique for proving robustness is adapted from methods commonly used for cryptographic protocols and our work illustrates the strengths and particularities of the ALEA style of reasoning about probabilistic programs.

3.4.2. Formalization of Languages, Semantics

- P. Herms, together with C. Marché and B. Monate (CEA List), has developed a certified VC generator, using Coq. The program for VC calculus and its specifications are both written in Coq, but the code is crafted so that it can be extracted automatically into a stand-alone executable. It is also designed in a way that allows the use of arbitrary first-order theorem provers to discharge the generated obligations [88]. On top of this generic VC generator, P. Herms developed a certified VC generator for C source code annotated using ACSL. This work is the main result of his PhD thesis [87].

[0] http://cvc4.cs.nyu.edu/web/

[0] http://www.cl.cam.ac.uk/~lp15/papers/Arith/

- A. Tafat and C. Marché have developed a certified VC generator using Why3 [97], [98]. The challenge was to formalize the operational semantics of an imperative language, and a corresponding weakest precondition calculus, without the possibility to use Coq advanced features such as dependent types or higher-order functions. The classical issues with local bindings, names and substitutions were solved by identifying appropriate lemmas. It was shown that Why3 can offer a significantly higher amount of proof automation compared to Coq.
- The full formalization of the JavaScript language specification, following the prose from the *ECMAScript Language Specification, version 5*, has been completed by the *JsCert* team [46], which includes A. Charguéraud from Toccata and 7 collaborators from Imperial College and Inria Rennes. For describing the 600+ evaluation rules, we have relied on a novel technique: the *pretty-big-step* semantics, which was developed by A. Charguéraud [7]. The formalization led to the discovery of bugs in the official standard, in the official test suites, and in all major browsers. It has raised the interest of several members of the ECMAScript standardization committee, and that of the developers of secure subsets for JavaScript.
- M. Clochard, C. Marché, and A. Paskevich have developed a general setting for developing programs involving binders, using Why3. This approach was successfully validated on two case studies: a verified implementation of untyped lambda-calculus and a verified tableaux-based theorem prover [64].
- M. Clochard, J.-C. Filliâtre, C. Marché, and A. Paskevich have developed a case study on the formalization of semantics of programming languages using Why3 [63]. This case study aims at illustrating recent improvements of Why3 regarding the support for higher-order logic features in the input logic of Why3, and how these are encoded into first-order logic, so that goals can be discharged by automated provers. This case study also illustrates how reasoning by induction can be done without need for interactive proofs, via the use of *lemma functions*.
- M. Clochard and L. Gondelman have developed a formalization of a simple compiler in Why3 [25]. It compiles a simple imperative language into assembler instructions for a stack machine. This case study was inspired by a similar example developed using Coq and interactive theorem proving. The aim is to improve significantly the degree of automation in the proofs. This is achieved by the formalization of a Hoare logic and a Weakest Precondition Calculus on assembly programs, so that the correctness of compilation is seen as a formal specification of the assembly instructions generated.
- S. Dumbrava and É. Contejean, with V. Benzaken (VALS team, at LRI) have proposed a *Coq* formalization of the relational data model which underlies relational database systems. More precisely, we have presented and formalized the data definition part of the model including integrity constraints. We have also modeled two different query language formalisms: relational algebra and conjunctive queries. Finally, we have presented logical query optimization and proved the main “database theorems”: algebraic equivalences, the homomorphism theorem and conjunctive query minimization [1].

3.4.3. Project-team Positioning

The objective of formalizing languages and algorithms is very general, and it is pursued by several Inria teams. One common trait is the use of the Coq proof assistant for this purpose: Pi.r2 (development of Coq itself and its meta-theory), Gallium (semantics and compilers of programming languages), Marelle (formalization of mathematics), SpecFun (real arithmetic), Celtique (formalization of static analyzers).

Other environments for the formalization of languages include

- ACL2 system [0]: an environment for writing programs with formal specifications in first-order logic based on a Lisp engine. The proofs are conducted using a prover based on the Boyer-Moore approach. It is a rather old system but still actively maintained and powerful, developed at University of Texas at Austin. It has a strong industrial impact.
- Isabelle environment [0]: both a proof assistant and an environment for developing pure applicative

[0] http://www.cs.utexas.edu/~moore/acl2/

programs. It is developed jointly at University of Cambridge, UK, Technische Universität München, Germany, and to some extent by the VALS team at LRI, Université Paris-Sud. It features highly automated tactics based on ATP systems (the Sledgehammer tool).

- The team "Trustworthy Systems" at NICTA in Australia [0] aims at developing highly trustable software applications. They developed a formally verified micro-kernel called seL4 [92], using a home-made layer to deal with C programs on top of the Isabelle prover.
- The PVS system [0] is an environment for both programming and proving (purely applicative) programs. It is developed at the Computer Science Laboratory of SRI international, California, USA. A major user of PVS is the team LFM [0] at NASA Langley, USA, for the certification of programs related to air traffic control.

In the Toccata team, we do not see these alternative environments as competitors, even though, for historical reasons, we are mainly using Coq. Indeed both Isabelle and PVS are available as back-ends of Why3.

3.5. Proof of Numerical Programs

Permanent researchers: S. Boldo, C. Marché, G. Melquiond

- Linked with objective 1 (Deductive Program Verification), the methodology for proving numerical C programs has been presented by S. Boldo in her habilitation [48] and as invited speaker [49]. An application is the formal verification of a numerical analysis program. S. Boldo, J.-C. Filliâtre, and G. Melquiond, with F. Clément and P. Weis (POMDAPI team, Inria Paris - Rocquencourt), and M. Mayero (LIPN), completed the formal proof of the second-order centered finite-difference scheme for the one-dimensional acoustic wave [51][5].
- Several challenging floating-point algorithms have been studied and proved. This includes an algorithm by Kahan for computing the area of a triangle: S. Boldo proved an improvement of its error bound and new investigations in case of underflow [47]. This includes investigations about quaternions. They should be of norm 1, but due to the round-off errors, a drift of this norm is observed over time. C. Marché determined a bound on this drift and formally proved it correct [99]. P. Roux formally verified an algorithm for checking that a matrix is semi-definite positive [19]. The challenge here is that testing semi-definiteness involves algebraic number computations, yet it needs to be implemented using only approximate floating-point operations.
- Because of compiler optimizations (or bugs), the floating-point semantics of a program might change once compiled, thus invalidating any property proved on the source code. We have investigated two ways to circumvent this issue, depending on whether the compiler is a black box. When it is, T. Nguyen has proposed to analyze the assembly code it generates and to verify it is correct [108]. On the contrary, S. Boldo and G. Melquiond (in collaboration with J.-H. Jourdan and X. Leroy) have added support for floating-point arithmetic to the CompCert compiler and formally proved that none of the transformations the compiler applies modify the floating-point semantics of the program [13] [53].
- Linked with objectives 2 (Automated Reasoning) and 3 (Formalization and Certification of Languages, Tools and Systems), G. Melquiond has implemented an efficient Coq library for floating-point arithmetic and proved its correctness in terms of operations on real numbers [102]. It serves as a basis for an interval arithmetic on which Taylor models have been formalized. É. Martin-Dorel and G. Melquiond have integrated these models into CoqInterval [18]. This Coq library is dedicated to automatically proving the approximation properties that occur when formally verifying the implementation of mathematical libraries (libm).

[0] http://isabelle.in.tum.de/

[0] http://ssrg.nicta.com.au/projects/TS/

[0] http://pvs.csl.sri.com/

[0] http://shemesh.larc.nasa.gov/fm/fm-main-team.html

- Double rounding occurs when the target precision of a floating-point computation is narrower than the working precision. In some situations, this phenomenon incurs a loss of accuracy. P. Roux has formally studied when it is innocuous for basic arithmetic operations [19]. É. Martin-Dorel and G. Melquiond (in collaboration with J.-M. Muller) have formally studied how it impacts algorithms used for error-free transformations [100]. These works were based on the Flocq formalization of floating-point arithmetic for Coq.
- By combining multi-precision arithmetic, interval arithmetic, and massively-parallel computations, G. Melquiond (in collaboration with G. Nowak and P. Zimmermann) has computed enough digits of the Masser-Gramain constant to invalidate a 30-year old conjecture about its closed form [103].

3.5.1. Project-team Positioning

This objective deals both with formal verification and floating-point arithmetic, which is quite uncommon. Therefore our competitors/peers are few. We may only cite the works by J. Duracz and M. Konečný, Aston University in Birmingham, UK.

The Inria team AriC (Grenoble - Rhône-Alpes) is closer to our research interests, but they are lacking manpower on the formal proof side; we have numerous collaborations with them. The Inria team Caramel (Nancy - Grand Est) also shares some research interests with us, though fewer; again, they do not work on the formal aspect of the verification; we have some occasional collaborations with them.

There are many formalization efforts from chip manufacturers, such as AMD (using the ACL2 proof assistant) and Intel (using the Forte proof assistants) but the algorithms they consider are quite different from the ones we study. The works on the topic of floating-point arithmetic from J. Harrison at Intel using HOL Light are really close to our research interests, but they seem to be discontinued.

A few deductive program verification teams are willing to extend their tools toward floating-point programs. This includes the KeY project and SPARK. We have an ongoing collaboration with the latter, in the context of the ProofInUSe project.

Deductive verification is not the only way to prove programs. Abstract interpretation is widely used, and several teams are interested in floating-point arithmetic. This includes the Inria team Antique (Paris - Rocquencourt) and a CEA List team, who have respectively developed the Astrée and Fluctuat tools. This approach targets a different class of numerical algorithms than the ones we are interested in.

Other people, especially from the SMT community (*cf* objective 2), are also interested in automatically proving formulas about floating-point numbers, notably at Oxford University. They are mainly focusing on pure floating-point arithmetic though and do not consider them as approximation of real numbers.

Finally, it can be noted that numerous teams are working on the verification of numerical programs, but assuming the computations are real rather than floating-point ones. This is out of the scope of this objective.

COMMANDS Project-Team

3. Research Program

3.1. Historical aspects

The roots of deterministic optimal control are the “classical” theory of the calculus of variations, illustrated by the work of Newton, Bernoulli, Euler, and Lagrange (whose famous multipliers were introduced in [65]), with improvements due to the “Chicago school”, Bliss [41] during the first part of the 20th century, and by the notion of relaxed problem and generalized solution (Young [73]).

Trajectory optimization really started with the spectacular achievement done by Pontryagin’s group [71] during the fifties, by stating, for general optimal control problems, nonlocal optimality conditions generalizing those of Weierstrass. This motivated the application to many industrial problems (see the classical books by Bryson and Ho [47], Leitmann [67], Lee and Markus [66], Ioffe and Tihomirov [62]). Since then, various theoretical achievements have been obtained by extending the results to nonsmooth problems, see Aubin [37], Clarke [48], Ekeland [55].

Dynamic programming was introduced and systematically studied by R. Bellman during the fifties. The HJB equation, whose solution is the value function of the (parameterized) optimal control problem, is a variant of the classical Hamilton-Jacobi equation of mechanics for the case of dynamics parameterized by a control variable. It may be viewed as a differential form of the dynamic programming principle. This nonlinear first-order PDE appears to be well-posed in the framework of *viscosity solutions* introduced by Crandall and Lions [50], [51], [49]. These tools also allow to perform the numerical analysis of discretization schemes. The theoretical contributions in this direction did not cease growing, see the books by Barles [39] and Bardi and Capuzzo-Dolcetta [38].

3.2. Trajectory optimization

The so-called *direct methods* consist in an optimization of the trajectory, after having discretized time, by a nonlinear programming solver that possibly takes into account the dynamic structure. So the two main problems are the choice of the discretization and the nonlinear programming algorithm. A third problem is the possibility of refinement of the discretization once after solving on a coarser grid.

In the *full discretization approach*, general Runge-Kutta schemes with different values of control for each inner step are used. This allows to obtain and control high orders of precision, see Hager [59], Bonnans [44]. In an interior-point algorithm context, controls can be eliminated and the resulting system of equation is easily solved due to its band structure. Discretization errors due to constraints are discussed in Dontchev et al. [54]. See also Malanowski et al. [68].

In the *indirect* approach, the control is eliminated thanks to Pontryagin’s maximum principle. One has then to solve the two-points boundary value problem (with differential variables state and costate) by a single or multiple shooting method. The questions are here the choice of a discretization scheme for the integration of the boundary value problem, of a (possibly globalized) Newton type algorithm for solving the resulting finite dimensional problem in $I\!R^n$ (n is the number of state variables), and a methodology for finding an initial point.

For state constrained problems or singular arcs, the formulation of the shooting function may be quite elaborate [42], [43], [36]. As initiated in [58], we focus more specifically on the handling of discontinuities, with ongoing work on the geometric integration aspects (Hamiltonian conservation).

3.3. Hamilton-Jacobi-Bellman approach

This approach consists in calculating the value function associated with the optimal control problem, and then synthesizing the feedback control and the optimal trajectory using Pontryagin's principle. The method has the great particular advantage of reaching directly the global optimum, which can be very interesting when the problem is not convex.

Characterization of the value function >From the dynamic programming principle, we derive a characterization of the value function as being a solution (in viscosity sense) of an Hamilton-Jacobi-Bellman equation, which is a nonlinear PDE of dimension equal to the number n of state variables. Since the pioneer works of Crandall and Lions [50], [51], [49], many theoretical contributions were carried out, allowing an understanding of the properties of the value function as well as of the set of admissible trajectories. However, there remains an important effort to provide for the development of effective and adapted numerical tools, mainly because of numerical complexity (complexity is exponential with respect to n).

Numerical approximation for continuous value function Several numerical schemes have been already studied to treat the case when the solution of the HJB equation (the value function) is continuous. Let us quote for example the Semi-Lagrangian methods [57], [56] studied by the team of M. Falcone (La Sapienza, Rome), the high order schemes WENO, ENO, Discrete galerkin introduced by S. Osher, C.-W. Shu, E. Harten [60], [61], [61], [69], and also the schemes on nonregular grids by R. Abgrall [35], [34]. All these schemes rely on finite differences or/and interpolation techniques which lead to numerical diffusions. Hence, the numerical solution is unsatisfying for long time approximations even in the continuous case.

One of the (nonmonotone) schemes for solving the HJB equation is based on the Ultrabee algorithm proposed, in the case of advection equation with constant velocity, by Roe [72] and recently revisited by Després-Lagoutière [53], [52]. The numerical results on several academic problems show the relevance of the antidiffusive schemes. However, the theoretical study of the convergence is a difficult question and is only partially done.

Optimal stochastic control problems occur when the dynamical system is uncertain. A decision typically has to be taken at each time, while realizations of future events are unknown (but some information is given on their distribution of probabilities). In particular, problems of economic nature deal with large uncertainties (on prices, production and demand). Specific examples are the portfolio selection problems in a market with risky and non-risky assets, super-replication with uncertain volatility, management of power resources (dams, gas). Air traffic control is another example of such problems.

Nonsmoothness of the value function. Sometimes the value function is smooth (e.g. in the case of Merton's portfolio problem, Oksendal [74]) and the associated HJB equation can be solved explicitly. Still, the value function is not smooth enough to satisfy the HJB equation in the classical sense. As for the deterministic case, the notion of viscosity solution provides a convenient framework for dealing with the lack of smoothness, see Pham [70], that happens also to be well adapted to the study of discretization errors for numerical discretization schemes [63], [40].

Numerical approximation for optimal stochastic control problems. The numerical discretization of second order HJB equations was the subject of several contributions. The book of Kushner-Dupuis [64] gives a complete synthesis on the Markov chain schemes (i.e Finite Differences, semi-Lagrangian, Finite Elements, ...). Here a main difficulty of these equations comes from the fact that the second order operator (i.e. the diffusion term) is not uniformly elliptic and can be degenerated. Moreover, the diffusion term (covariance matrix) may change direction at any space point and at any time (this matrix is associated the dynamics volatility).

For solving stochastic control problems, we studied the so-called Generalized Finite Differences (GFD), that allow to choose at any node, the stencil approximating the diffusion matrix up to a certain threshold [46]. Determining the stencil and the associated coefficients boils down to a quadratic program to be solved at each point of the grid, and for each control. This is definitely expensive, with the exception of special structures where the coefficients can be computed at low cost. For two dimensional systems, we designed a (very) fast algorithm for computing the coefficients of the GFD scheme, based on the Stern-Brocot tree [45].

DEFI Project-Team

3. Research Program

3.1. Research Program

The research activity of our team is dedicated to the design, analysis and implementation of efficient numerical methods to solve inverse and shape/topological optimization problems in connection with wave imaging, structural design, non-destructive testing and medical imaging modalities. We are particularly interested in the development of fast methods that are suited for real-time applications and/or large scale problems. These goals require to work on both the physical and the mathematical models involved and indeed a solid expertise in related numerical algorithms.

This section intends to give a general overview of our research interests and themes. We choose to present them through the specific academic example of inverse scattering problems (from inhomogeneities), which is representative of foreseen developments on both inversion and (topological) optimization methods. The practical problem would be to identify an inclusion from measurements of diffracted waves that result from the interaction of the sought inclusion with some (incident) waves sent into the probed medium. Typical applications include biomedical imaging where using micro-waves one would like to probe the presence of pathological cells, or imaging of urban infrastructures where using ground penetrating radars (GPR) one is interested in finding the location of buried facilities such as pipelines or waste deposits. This kind of applications requires in particular fast and reliable algorithms.

By “imaging” we shall refer to the inverse problem where the concern is only the location and the shape of the inclusion, while “identification” may also indicate getting informations on the inclusion physical parameters.

Both problems (imaging and identification) are non linear and ill-posed (lack of stability with respect to measurements errors if some careful constrains are not added). Moreover, the unique determination of the geometry or the coefficients is not guaranteed in general if sufficient measurements are not available. As an example, in the case of anisotropic inclusions, one can show that an appropriate set of data uniquely determine the geometry but not the material properties.

These theoretical considerations (uniqueness, stability) are not only important in understanding the mathematical properties of the inverse problem, but also guide the choice of appropriate numerical strategies (which information can be stably reconstructed) and also the design of appropriate regularization techniques. Moreover, uniqueness proofs are in general constructive proofs, i.e. they implicitly contain a numerical algorithm to solve the inverse problem, hence their importance for practical applications. The sampling methods introduced below are one example of such algorithms.

A large part of our research activity is dedicated to numerical methods applied to the first type of inverse problems, where only the geometrical information is sought. In its general setting the inverse problem is very challenging and no method can provide a universal satisfactory solution to it (regarding the balance cost-precision-stability). This is why in the majority of the practically employed algorithms, some simplification of the underlying mathematical model is used, according to the specific configuration of the imaging experiment. The most popular ones are geometric optics (the Kirchhoff approximation) for high frequencies and weak scattering (the Born approximation) for small contrasts or small obstacles. They actually give full satisfaction for a wide range of applications as attested by the large success of existing imaging devices (radar, sonar, ultrasound, X-ray tomography, etc.), that rely on one of these approximations.

Generally speaking, the used simplifications result in a linearization of the inverse problem and therefore are usually valid only if the latter is weakly non-linear. The development of these simplified models and the improvement of their efficiency is still a very active research area. With that perspective we are particularly interested in deriving and studying higher order asymptotic models associated with small geometrical parameters such as: small obstacles, thin coatings, wires, periodic media, ... Higher order models usually introduce some non linearity in the inverse problem, but are in principle easier to handle from the numerical point of view than in the case of the exact model.

A larger part of our research activity is dedicated to algorithms that avoid the use of such approximations and that are efficient where classical approaches fail: i.e. roughly speaking when the non linearity of the inverse problem is sufficiently strong. This type of configuration is motivated by the applications mentioned below, and occurs as soon as the geometry of the unknown media generates non negligible multiple scattering effects (multiply-connected and closely spaces obstacles) or when the used frequency is in the so-called resonant region (wave-length comparable to the size of the sought medium). It is therefore much more difficult to deal with and requires new approaches. Our ideas to tackle this problem will be motivated and inspired by recent advances in shape and topological optimization methods and also the introduction of novel classes of imaging algorithms, so-called sampling methods.

The sampling methods are fast imaging solvers adapted to multi-static data (multiple receiver-transmitter pairs) at a fixed frequency. Even if they do not use any linearization the forward model, they rely on computing the solutions to a set of linear problems of small size, that can be performed in a completely parallel procedure. Our team has already a solid expertise in these methods applied to electromagnetic 3-D problems. The success of such approaches was their ability to provide a relatively quick algorithm for solving 3-D problems without any need for a priori knowledge on the physical parameters of the targets. These algorithms solve only the imaging problem, in the sense that only the geometrical information is provided.

Despite the large efforts already spent in the development of this type of methods, either from the algorithmic point of view or the theoretical one, numerous questions are still open. These attractive new algorithms also suffer from the lack of experimental validations, due to their relatively recent introduction. We also would like to invest on this side by developing collaborations with engineering research groups that have experimental facilities. From the practical point of view, the most potential limitation of sampling methods would be the need of a large amount of data to achieve a reasonable accuracy. On the other hand, optimization methods do not suffer from this constrain but they require good initial guess to ensure convergence and reduce the number of iterations. Therefore it seems natural to try to combine the two class of methods in order to calibrate the balance between cost and precision.

Among various shape optimization methods, the Level Set method seems to be particularly suited for such a coupling. First, because it shares similar mechanism as sampling methods: the geometry is captured as a level set of an “indicator function” computed on a cartesian grid. Second, because the two methods do not require any a priori knowledge on the topology of the sought geometry. Beyond the choice of a particular method, the main question would be to define in which way the coupling can be achieved. Obvious strategies consist in using one method to pre-process (initialization) or post-process (find the level set) the other. But one can also think of more elaborate ones, where for instance a sampling method can be used to optimize the choice of the incident wave at each iteration step.The latter point is closely related to the design of so called “focusing incident waves” (which are for instance the basis of applications of the time-reversal principle). In the frequency regime, these incident waves can be constructed from the eigenvalue decomposition of the data operator used by sampling methods. The theoretical and numerical investigations of these aspects are still not completely understood for electromagnetic or elastodynamic problems.

Other topological optimization methods, like the homogenization method or the topological gradient method, can also be used, each one provides particular advantages in specific configurations. It is evident that the development of these methods is very suited to inverse problems and provide substantial advantage compared to classical shape optimization methods based on boundary variation. Their applications to inverse problems has not been fully investigated. The efficiency of these optimization methods can also be increased for adequate asymptotic configurations. For instance small amplitude homogenization method can be used as an efficient relaxation method for the inverse problem in the presence of small contrasts. On the other hand, the topological gradient method has shown to perform well in localizing small inclusions with only one iteration.

A broader perspective would be the extension of the above mentioned techniques to time-dependent cases. Taking into account data in time domain is important for many practical applications, such as imaging in cluttered media, the design of absorbing coatings or also crash worthiness in the case of structural design.

For the identification problem, one would like to also have information on the physical properties of the targets. Of course optimization methods is a tool of choice for these problems. However, in some applications only a qualitative information is needed and obtaining it in a cheaper way can be performed using asymptotic theories combined with sampling methods. We also refer here to the use of so called transmission eigenvalues as qualitative indicators for non destructive testing of dielectrics.

We are also interested in parameter identification problems arising in diffusion-type problems. Our research here is mostly motivated by applications to the imaging of biological tissues with the technique of Diffusion Magnetic Resonance Imaging (DMRI). Roughly speaking DMRI gives a measure of the average distance travelled by water molecules in a certain medium and can give useful information on cellular structure and structural change when the medium is biological tissue. In particular, we would like to infer from DMRI measurements changes in the cellular volume fraction occurring upon various physiological or pathological conditions as well as the average cell size in the case of tumor imaging. The main challenges here are 1) correctly model measured signals using diffusive-type time-dependent PDEs 2) numerically handle the complexity of the tissues 3) use the first two to identify physically relevant parameters from measurements. For the last point we are particularly interested in constructing reduced models of the multiple-compartment Bloch-Torrey partial differential equation using homogenization methods.

DISCO Project-Team

3. Research Program

3.1. Modeling of complex environment

We want to model phenomena such as a temporary loss of connection (e.g. synchronisation of the movements through haptic interfaces), a nonhomogeneous environment (e.g. case of cryogenic systems) or the presence of the human factor in the control loop (e.g. grid systems) but also problems involved with technological constraints (e.g. range of the sensors). The mathematical models concerned include integro-differential, partial differential equations, algebraic inequalities with the presence of several time scales, whose variables and/or parameters must satisfy certain constraints (for instance, positivity).

3.2. Analysis of interconnected systems

- Algebraic analysis of linear systems

  Study of the structural properties of linear differential time-delay systems and linear infinite-dimensional systems (e.g. invariants, controllability, observability, flatness, reductions, decomposition, decoupling, equivalences) by means of constructive algebra, module theory, homological algebra, algebraic analysis and symbolic computation [8], [9], [79], [94], [80], [81].

- Robust stability of linear systems

  Within an interconnection context, lots of phenomena are modelled directly or after an approximation by delay systems. These systems might have fixed delays, time-varying delays, distributed delays ...

  For various infinite-dimensional systems, particularly delay and fractional systems, input-output and time-domain methods are jointly developed in the team to characterize stability. This research is developed at four levels: analytic approaches (H_∞-stability, BIBO-stablity, robust stability, robustness metrics) [1], [2], [5], [6], symbolic computation approaches (SOS methods are used for determining easy-to-check conditions which guarantee that the poles of a given linear system are not in the closed right half-plane, certified CAD techniques), numerical approaches (root-loci, continuation methods) and by means of softwares developed in the team [5], [6].

- Robustness/fragility of biological systems

  Deterministic biological models describing, for instance, species interactions, are frequently composed of equations with important disturbances and poorly known parameters. To evaluate the impact of the uncertainties, we use the techniques of designing of global strict Lyapunov functions or functional developed in the team.

  However, for other biological systems, the notion of robustness may be different and this question is still in its infancy (see, e.g. [90]). Unlike engineering problems where a major issue is to maintain stability in the presence of disturbances, a main issue here is to maintain the system response in the presence of disturbances. For instance, a biological network is required to keep its functioning in case of a failure of one of the nodes in the network. The team, which has a strong expertise in robustness for engineering problems, aims at contributing at the develpment of new robustness metrics in this biological context.

3.3. Stabilization of interconnected systems

- Linear systems: Analytic and algebraic approaches are considered for infinite-dimensional linear systems studied within the input-output framework.

  In the recent years, the Youla-Kučera parametrization (which gives the set of all stabilizing controllers of a system in terms of its coprime factorizations) has been the cornerstone of the success of the H_∞-control since this parametrization allows one to rewrite the problem of finding the optimal stabilizing controllers for a certain norm such as H_∞ or H_2 as affine, and thus, convex problem.

  A central issue studied in the team is the computation of such factorizations for a given infinite-dimensional linear system as well as establishing the links between stabilizability of a system for a certain norm and the existence of coprime factorizations for this system. These questions are fundamental for robust stabilization problems [1], [2], [8], [9].

  We also consider simultaneous stabilization since it plays an important role in the study of reliable stabilization, i.e. in the design of controllers which stabilize a finite family of plants describing a system during normal operating conditions and various failed modes (e.g. loss of sensors or actuators, changes in operating points) [9]. Moreover, we investigate strongly stabilizable systems [9], namely systems which can be stabilized by stable controllers, since they have a good ability to track reference inputs and, in practice, engineers are reluctant to use unstable controllers especially when the system is stable.

- Nonlinear systems

  The project aims at developing robust stabilization theory and methods for important classes of nonlinear systems that ensure good controllerperformance under uncertainty and time delays. The main techniques include techniques called backstepping and forwarding, contructions of strict Lyapunov functions through so-called "strictification" approaches [3] and construction of Lyapunov-Krasovskii functionals [4], [5], [6].

- Predictive control

  For highly complex systems described in the time-domain and which are submitted to constraints, predictive control seems to be well-adapted. This model based control method (MPC: Model Predictive Control) is founded on the determination of an optimal control sequence over a receding horizon. Due to its formulation in the time-domain, it is an effective tool for handling constraints and uncertainties which can be explicitly taken into account in the synthesis procedure [7]. The team considers how mutiparametric optimization can help to reduce the computational load of this method, allowing its effective use on real world constrained problems.

  The team also investigates stochastic optimization methods such as genetic algorithm, particle swarm optimization or ant colony [10] as they can be used to optimize any criterion and constraint whatever their mathematical structure is. The developed methodologies can be used by non specialists.

3.4. Synthesis of reduced complexity controllers

- PID controllers

  Even though the synthesis of control laws of a given complexity is not a new problem, it is still open, even for finite-dimensional linear systems. Our purpose is to search for good families of “simple” (e.g. low order) controllers for infinite-dimensional dynamical systems. Within our approach, PID candidates are first considered in the team [2], [93].

- Predictive control

  The synthesis of predictive control laws is concerned with the solution of multiparametric optimization problems. Reduced order controller constraints can be viewed as non convex constraints in the synthesis procedure. Such constraints can be taken into account with stochastic algorithms.

Finally, the development of algorithms based on both symbolic computation and numerical methods, and their implementations in dedicated Scilab/Matlab/Maple toolboxes are important issues in the project.

GECO Project-Team

3. Research Program

3.1. Geometric control theory

The main research topic of the project-team will be **geometric control**, with a special focus on **control design**. The application areas that we target are control of quantum mechanical systems, neurogeometry and switched systems.

Geometric control theory provides a viewpoint and several tools, issued in particular from differential geometry, to tackle typical questions arising in the control framework: controllability, observability, stabilization, optimal control... [30], [64] The geometric control approach is particularly well suited for systems involving nonlinear and nonholonomic phenomena. We recall that nonholonomicity refers to the property of a velocity constraint that is not equivalent to a state constraint.

The expression **control design** refers here to all phases of the construction of a control law, in a mainly open-loop perspective: modeling, controllability analysis, output tracking, motion planning, simultaneous control algorithms, tracking algorithms, performance comparisons for control and tracking algorithms, simulation and implementation.

We recall that

- **controllability** denotes the property of a system for which any two states can be connected by a trajectory corresponding to an admissible control law ;
- **output tracking** refers to a control strategy aiming at keeping the value of some functions of the state arbitrarily close to a prescribed time-dependent profile. A typical example is **configuration tracking** for a mechanical system, in which the controls act as forces and one prescribes the position variables along the trajectory, while the evolution of the momenta is free. One can think for instance at the lateral movement of a car-like vehicle: even if such a movement is unfeasible, it can be tracked with arbitrary precision by applying a suitable control strategy;
- **motion planning** is the expression usually denoting the algorithmic strategy for selecting one control law steering the system from a given initial state to an attainable final one;
- **simultaneous control** concerns algorithms that aim at driving the system from two different initial conditions, with the same control law and over the same time interval, towards two given final states (one can think, for instance, at some control action on a fluid whose goal is to steer simultaneously two floating bodies.) Clearly, the study of which pairs (or n-uples) of states can be simultaneously connected thanks to an admissible control requires an additional controllability analysis with respect to the plain controllability mentioned above.

At the core of control design is then the notion of motion planning. Among the motion planning methods, a preeminent role is played by those based on the Lie algebra associated with the control system ([84], [71], [77]), those exploiting the possible flatness of the system ([58]) and those based on the continuation method ([96]). Optimal control is clearly another method for choosing a control law connecting two states, although it generally introduces new computational and theoretical difficulties.

Control systems with special structure, which are very important for applications are those for which the controls appear linearly. When the controls are not bounded, this means that the admissible velocities form a distribution in the tangent bundle to the state manifold. If the distribution is equipped with a smoothly varying norm (representing a cost of the control), the resulting geometrical structure is called *sub-Riemannian*. Sub-Riemannian geometry thus appears as the underlying geometry of the nonholonomic control systems, playing the same role as Euclidean geometry for linear systems. As such, its study is fundamental for control design. Moreover its importance goes far beyond control theory and is an active field of research both in differential geometry ([83]), geometric measure theory ([59], [34]) and hypoelliptic operator theory ([46]).

Other important classes of control systems are those modeling mechanical systems. The dynamics are naturally defined on the tangent or cotangent bundle of the configuration manifold, they have Lagrangian or Hamiltonian structure, and the controls act as forces. When the controls appear linearly, the resulting model can be seen somehow as a second-order sub-Riemannian structure (see [51]).

The control design topics presented above naturally extend to the case of distributed parameter control systems. The geometric approach to control systems governed by partial differential equations is a novel subject with great potential. It could complement purely analytical and numerical approaches, thanks to its more dynamical, qualitative and intrinsic point of view. An interesting example of this approach is the paper [31] about the controllability of Navier–Stokes equation by low forcing modes.

Maxplus Team

3. Research Program

3.1. L'algèbre max-plus/Max-plus algebra

Le semi-corps *max-plus* est l'ensemble $\mathbb{R} \cup \{-\infty\}$, muni de l'addition $(a, b) \mapsto a \oplus b = \max(a, b)$ et de la multiplication $(a, b) \mapsto a \otimes b = a + b$. Cette structure algébrique diffère des structures de corps classiques par le fait que l'addition n'est pas une loi de groupe, mais est idempotente: $a \oplus a = a$. On rencontre parfois des variantes de cette structure: par exemple, le semi-corps *min-plus* est l'ensemble $\mathbb{R} \cup \{+\infty\}$ muni des lois $a \oplus b = \min(a, b)$ et $a \otimes b = a + b$, et le semi-anneau *tropical* est l'ensemble $\mathbb{N} \cup \{+\infty\}$ munis des mêmes lois. L'on peut se poser la question de généraliser les constructions de l'algèbre et de l'analyse classique, qui reposent pour une bonne part sur des anneaux ou des corps tels que $\mathbb{Z}$ ou $\mathbb{R}$, au cas de semi-anneaux de type max-plus: tel est l'objet de ce qu'on appelle un peu familièrement "l'algèbre max-plus".

Il est impossible ici de donner une vue complète du domaine. Nous nous bornerons à indiquer quelques références bibliographiques. L'intérêt pour les structures de type max-plus est contemporain de la naissance de la théorie des treillis [103]. Depuis, les structures de type max-plus ont été développées indépendamment par plusieurs écoles, en relation avec plusieurs domaines. Les motivations venant de la Recherche Opérationnelle (programmation dynamique, problèmes de plus court chemin, problèmes d'ordonnancement, optimisation discrète) ont été centrales dans le développement du domaine [92], [121], [166], [169], [170]. Les semi-anneaux de type max-plus sont bien sûr reliés aux algèbres de Boole [79]. L'algèbre max-plus apparaît de manière naturelle en contrôle optimal et dans la théorie des équations aux dérivées partielles d'Hamilton-Jacobi [156], [155], [144], [129], [118], [159], [138], [119], [106], [52]. Elle apparaît aussi en analyse asymptotique (asymptotiques de type WKB [143], [144], [129], grandes déviations [153], asymptotiques à température nulle en physique statistique [81]), puisque l'algèbre max-plus apparaît comme limite de l'algèbre usuelle. La théorie des opérateurs linéaires max-plus peut être vue comme faisant partie de la théorie des opérateurs de Perron-Frobenius non-linéaires, ou de la théorie des applications contractantes ou monotones sur les cônes [130], [149], [141], [69], laquelle a de nombreuse motivations, telles l'économie mathématique [147], et la théorie des jeux [157], [39]. Dans la communauté des systèmes à événements discrets, l'algèbre max-plus a été beaucoup étudiée parce qu'elle permet de représenter de manière linéaire les phénomènes de synchronisation, lesquels déterminent le comportement temporel de systèmes de production ou de réseaux, voir [6]. Parmi les développements récents du domaine, on peut citer le calcul des réseaux [80], [134], qui permet de calculer des bornes pire des cas de certaines mesures de qualité de service. En informatique théorique, l'algèbre max-plus (ou plutôt le semi-anneau tropical) a joué un rôle décisif dans la résolution de problèmes de décision en théorie des automates [161], [124], [162], [131], [151]. Notons finalement, pour information, que l'algèbre max-plus est apparue récemment en géométrie algébrique [117], [165], [146], [164] et en théorie des représentations [107], [72], sous les noms de géométrie et combinatoire tropicales.

Nous décrivons maintenant de manière plus détaillée les sujets qui relèvent directement des intérêts du projet, comme la commande optimale, les asymptotiques, et les systèmes à événements discrets.

English version

The *max-plus* semifield is the set $\mathbb{R} \cup \{-\infty\}$, equipped with the addition $(a, b) \mapsto a \oplus b = \max(a, b)$ and the multiplication $(a, b) \mapsto a \otimes b = a + b$. This algebraic structure differs from classical structures, like fields, in that addition is idempotent: $a \oplus a = a$. Several variants have appeared in the literature: for instance, the *min-plus* semifield is the set $\mathbb{R} \cup \{+\infty\}$ equipped with the laws $a \oplus b = \min(a, b)$ and $a \otimes b = a + b$, and the *tropical* semiring is the set $\mathbb{N} \cup \{+\infty\}$ equipped with the same laws. One can ask the question of extending to max-plus type structures the classical constructions and results of algebra and analysis: this is what is often called in a wide sense "max-plus algebra" or "tropical algebra".

It is impossible to give in this short space a fair view of the field. Let us, however, give a few references. The interest in max-plus type structures is contemporaneous with the early developments of lattice theory [103]. Since that time, max-plus structures have been developed independently by several schools, in relation with several fields. Motivations from Operations Research (dynamic programming, shortest path problems, scheduling problems, discrete optimisation) were central in the development of the field [92], [121], [166], [169], [170]. Of course, max-plus type semirings are related to Boolean algebras [79]. Max-plus algebras arises naturally in optimal control and in the theory of Hamilton-Jacobi partial differential equations [156], [155], [144], [129], [118], [159], [138], [119], [106], [52]. It arises in asymptotic analysis (WKB asymptotics [143], [144], [129], large deviation asymptotics [153], or zero temperature asymptotics in statistical physics [81]), since max-plus algebra appears as a limit of the usual algebra. The theory of max-plus linear operators may be thought of as a part of the non-linear Perron-Frobenius theory, or of the theory of nonexpansive or monotone operators on cones [130], [149], [141], [69], a theory with numerous motivations, including mathematical economy [147] and game theory [157], [39]. In the discrete event systems community, max-plus algebra has been much studied since it allows one to represent linearly the synchronisation phenomena which determine the time behaviour of manufacturing systems and networks, see [6]. Recent developments include the network calculus of [80], [134] which allows one to compute worst case bounds for certain measures of quality of service. In theoretical computer science, max-plus algebra (or rather, the tropical semiring) played a key role in the solution of decision problems in automata theory [161], [124], [162], [131], [151]. We finally note for information that max-plus algebra has recently arisen in algebraic geometry [117], [165], [146], [164] and in representation theory [107], [72], under the names of tropical geometry and combinatorics.

We now describe in more details some parts of the subject directly related to our interests, like optimal control, asymptotics, and discrete event systems.

3.2. Algèbre max-plus, programmation dynamique, et commande optimale/Max-plus algebra, dynamic programming, and optimal control

L’exemple le plus simple d’un problème conduisant à une équation min-plus linéaire est le problème classique du plus court chemin. Considérons un graphe dont les nœuds sont numérotés de 1 à n et dont le coût de l’arc allant du nœud i au nœud j est noté $M_{ij} \in \mathbb{R} \cup \{+\infty\}$. Le coût minimal d’un chemin de longueur k, allant de i à j, est donné par la quantité:

$$v_{ij}(k) = \min_{\ell :\ \ell_0=i,\ \ell_k=j} \sum_{r=0}^{k-1} M_{\ell_r \ell_{r+1}} \quad , \tag{1}$$

où le minimum est pris sur tous les chemins $\ell = (\ell_0, ..., \ell_k)$ de longueur k, de nœud initial $\ell_0 = i$ et de nœud final $\ell_k = j$. L’équation classique de la programmation dynamique s’écrit:

$$v_{ij}(k) = \min_{1 \le s \le n} (M_{is} + v_{sj}(k-1)) \quad . \tag{2}$$

On reconnaît ainsi une équation linéaire min-plus :

$$v(k) = Mv(k-1) \quad , \tag{3}$$

où on note par la concaténation le produit matriciel induit par la structure de l’algèbre min-plus. Le classique *problème de Lagrange* du calcul des variations,

$$v(x,T) = \inf_{X(\cdot),\ \ X(0)=x} \int_0^T L(X(t), \dot{X}(t))\mathrm{d}t + \phi(X(T)) \quad , \tag{4}$$

où $X(t) \in \mathbb{R}^n$, pour $0 \leq t \leq T$, et $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est le Lagrangien, peut être vu comme une version continue de (1), ce qui permet de voir l'équation d'Hamilton-Jacobi que vérifie v,

$$v(\cdot, 0) = \phi, \quad \frac{\partial v}{\partial T} + H(x, \frac{\partial v}{\partial x}) = 0, \qquad H(x, p) = \sup_{y \in \mathbb{R}^n} (-p \cdot y - L(x, y)) \ , \tag{5}$$

comme une équation min-plus linéaire. En particulier, les solutions de (5) vérifient un principe de superposition min-plus: si v et w sont deux solutions, et si $\lambda, \mu \in \mathbb{R}$, $\inf(\lambda + v, \mu + w)$ est encore solution de (5). Ce point de vue, inauguré par Maslov, a conduit au développement de l'école d'Analyse Idempotente (voir [144], [129], [138]).

La présence d'une structure algébrique sous-jacente permet de voir les solutions stationnaires de (2) et (5) comme des vecteurs propres de la matrice M ou du semi-groupe d'évolution de l'équation d'Hamilton-Jacobi. La valeur propre associée fournit le coût moyen par unité de temps (coût ergodique). La représentation des vecteurs propres (voir [156], [166], [92], [120], [86], [68], [6] pour la dimension finie, et [144], [129] pour la dimension infinie) est intimement liée au théorème de l'autoroute qui décrit les trajectoires optimales quand la durée ou la longueur des chemins tend vers l'infini. Pour l'équation d'Hamilton-Jacobi, des résultats reliés sont apparus récemment en théorie d'"Aubry-Mather" [106].

English version

The most elementary example of a problem leading to a min-plus linear equation is the classical shortest path problem. Consider a graph with nodes $1, ..., n$, and let $M_{ij} \in \mathbb{R} \cup \{+\infty\}$ denote the cost of the arc from node i to node j. The minimal cost of a path of a given length, k, from i to j, is given by (1), where the minimum is taken over all paths $\ell = (\ell_0, ..., \ell_k)$ of length k, with initial node $\ell_0 = i$ and final node $\ell_k = j$. The classical dynamic programming equation can be written as in (2). We recognise the min-plus linear equation (3), where concatenation denotes the matrix product induced by the min-plus algebraic structure. The classical *Lagrange problem* of calculus of variations, given by (4) where $X(t) \in \mathbb{R}^n$, for $0 \leq t \leq T$, and $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the Lagrangian, may be thought of as a continuous version of (1), which allows us to see the Hamilton-Jacobi equation (5) satisfied by v, as a min-plus linear equation. In particular, the solutions of (5) satisfy a min-plus superposition principle: if v and w are two solutions, and if $\lambda, \mu \in \mathbb{R}$, then $\inf(\lambda + v, \mu + w)$ is also a solution of (5). This point of view, due to Maslov, led to the developpement of the school of Idempotent Analysis (see [144], [129], [138]).

The underlying algebraic structure allows one to see stationnary solutions of (2) and (5) as eigenvectors of the matrix M or of the evolution semigroup of the Hamilton-Jacobi equation. The associated eigenvalue gives the average cost per time unit (ergodic cost). The representation of eigenvectors (see [156], [166], [120], [86], [92], [68], [6] for the finite dimension case, and [144], [129] for the infinite dimension case) is intimately related to turnpike theorems, which describe optimal trajectories as the horizon, or path length, tends to infinity. For the Hamilton-Jacobi equation, related results have appeared recently in the "Aubry-Mather" theory [106].

3.3. Applications monotones et théorie de Perron-Frobenius non-linéaire, ou l'approche opératorielle du contrôle optimal et des jeux/Monotone maps and non-linear Perron-Frobenius theory, or the operator approach to optimal control and games

On sait depuis le tout début des travaux en décision markovienne que les opérateurs de la programmation dynamique f de problèmes de contrôle optimal ou de jeux (à somme nulle et deux joueurs), avec critère additif, ont les propriétés suivantes :

$$\begin{array}{llcl} \text{monotonie}/\textit{monotonicity} & x \leq y & \Rightarrow & f(x) \leq f(y) \ , \\ \text{contraction}/\textit{nonexpansiveness} & \|f(x) - f(y)\|_\infty & \leq & \|x - y\|_\infty \ . \end{array} \tag{6}$$

Ici, l'opérateur f est une application d'un certain espace de fonctions à valeurs réelles dans lui-même, $\leq$ désigne l'ordre partiel usuel, et $\|\cdot\|_\infty$ désigne la norme sup. Dans le cas le plus simple, l'ensemble des états est $\{1,...,n\}$ et f est une application de $\mathbb{R}^n$ dans lui-même. Les applications monotones qui sont contractantes pour la norme du sup peuvent être vues comme des généralisations non-linéaires des matrices sous-stochastiques. Une sous-classe utile, généralisant les matrices stochastiques, est formée des applications qui sont monotones et commutent avec l'addition d'une constante [91] (celles ci sont parfois appelées fonctions topicales). Les problèmes de programmation dynamique peuvent être traduits en termes d'opérateurs : l'équation de la programmation dynamique d'un problème de commande optimale à horizon fini s'écrit en effet $x(k) = f(x(k-1))$, où $x(k)$ est la fonction valeur en horizon k et $x(0)$ est donné; la fonction valeur y d'un problème à horizon infini (y compris le cas d'un problème d'arrêt optimal) vérifie $y = f(y)$; la fonction valeur z d'un problème avec facteur d'actualisation $0 < \alpha < 1$ vérifie $z = f(\alpha z)$, etc. Ce point de vue abstrait a été très fructueux, voir par exemple [39]. Il permet d'inclure la programmation dynamique dans la perspective plus large de la théorie de Perron-Frobenius non-linéaire, qui, depuis l'extension du théorème de Perron-Frobenius par Krein et Rutman, traite des applications non linéaires sur des cônes vérifiant des conditions de monotonie, de contraction ou d'homogénéité. Les problèmes auxquels on s'intéresse typiquement sont la structure de l'ensemble des points fixes de f, le comportement asymptotique de f^k, en particulier l'existence de la limite de $f^k(x)/k$ lorsque k tends vers l'infini (afin d'obtenir le coût ergodique d'un problème de contrôle optimal ou de jeux), l'asymptotique plus précise de f^k, à une normalisation près (afin d'obtenir le comportement précis de l'itération sur les valeurs), etc. Nous renvoyons le lecteur à [149] pour un panorama. Signalons que dans [111],[7], des algorithmes inspirés de l'algorithme classique d'itérations sur les politiques du contrôle stochastique ont pu être introduits dans le cas des opérateurs monotones contractants généraux, en utilisant des résultats de structure de l'ensemble des points fixes de ces opérateurs. Les applications de la théorie des applications monotones contractantes ne se limitent pas au contrôle optimal et aux jeux. En particulier, on utilise la même classe d'applications dans la modélisation des systèmes à événements discrets, voir le §3.5 ci-dessous, et une classe semblable d'applications en analyse statique de programmes, voir §4.4 ci-dessous.

English version

Since the very beginning of Markov decision theory, it has been observed that dynamic programming operators f arising in optimal control or (zero-sum, two player) game problems have Properties (6). Here, the operator f is a self-map of a certain space of real valued functions, equipped with the standard ordering $\leq$ and with the sup-norm $\|\cdot\|_\infty$. In the simplest case, the set of states is $\{1,...,n\}$, and f is a self-map of $\mathbb{R}^n$. Monotone maps that are nonexpansive in the sup norm may be thought of as nonlinear generalisations of substochastic matrices. A useful subclass, which generalises stochastic matrices, consists of those maps which are monotone and commute with the addition of a constant [91] (these maps are sometimes called topical functions). Dynamic programming problems can be translated in operator terms: the dynamic programming equation for a finite horizon problem can be written as $x(k) = f(x(k-1))$, where $x(k)$ is the value function in horizon k and $x(0)$ is given; the value function y of a problem with an infinite horizon (including the case of optimal stopping) satisfies $y = f(y)$; the value function z of a problem with discount factor $0 < \alpha < 1$ satisfies $z = f(\alpha z)$, etc. This abstract point of view has been very fruitful, see for instance [39]. It allows one to put dynamic programming in the wider perspective of nonlinear Perron-Frobenius theory, which, after the extension of the Perron-Frobenius theorem by Krein and Rutman, studies non-linear self-maps of cones, satisfying various monotonicity, nonexpansiveness, and homogeneity conditions. Typical problems of interests are the structure of the fixed point set of f, the asymptotic behaviour of f^k, including the existence of the limit of $f^k(x)/k$ as k tends to infinity (which yields the ergodic cost in control or games problems), the finer asymptotic behaviour of f^k, possibly up to a normalisation (which yields precise results on value iteration), etc. We shall not attempt to survey this theory here, and will only refer the reader to [149] for more background. In [111],[7], algorithms inspired from the classical policy iterations algorithm of stochastic control have been introduced for general monotone nonexpansive operators, using structural results for the fixed point set of these operators. Applications of monotone or nonexpansive maps are not limited to optimal control and game theory. In particular, we also use the same class of maps as models of discrete event dynamics systems, see §3.5 below, and we shall see in §4.4 that related classes of maps are useful in the static analysis of computer programs.

3.4. Processus de Bellman/Bellman processes

Un autre point de vue sur la commande optimale est la théorie des *processus de Bellman* [155], [96], [95], [52],[1], qui fournit un analogue max-plus de la théorie des probabilités. Cette théorie a été développée à partir de la notion de *mesure idempotente* introduite par Maslov [143]. Elle établit une correspondance entre probabilités et optimisation, dans laquelle les variables aléatoires deviennent des variables de coût (qui permettent de paramétriser les problèmes d'optimisation), la notion d'espérance conditionnelle est remplacée par celle de coût conditionnel (pris sur un ensemble de solutions faisables), la propriété de Markov correspond au principe de la programmation dynamique de Bellman, et la convergence faible à une convergence de type épigraphe. Les théorèmes limites pour les processus de Bellman (loi des grands nombres, théorème de la limite centrale, lois stables) fournissent des résultats asymptotiques en commande optimale. Ces résultats généraux permettent en particulier de comprendre qualitativement les difficultés d'approximation des solutions d'équations d'Hamilton-Jacobi retrouvés en particulier dans le travail de thèse d'Asma Lakhoua [132], [48].

English version

Another point of view on optimal control is the theory of *Bellman processes* [155], [96], [95], [52], [1] which provides a max-plus analogue of probability theory, relying on the theory of *idempotent measures* due to Maslov [143]. This establishes a correspondence between probability and optimisation, in which random variables become cost variables (which allow to parametrise optimisation problems), the notion of conditional expectation is replaced by a notion of conditional cost (taken over a subset of feasible solutions), the Markov property corresponds to the Bellman's dynamic programming principle, and weak convergence corresponds to an epigraph-type convergence. Limit theorems for Bellman processes (law of large numbers, central limit theorems, stable laws) yield asymptotic results in optimal control. Such general results help in particular to understand qualitatively the difficulty of approximation of Hamilton-Jacobi equations found again in particular in the PhD thesis work of Asma Lakhoua [132], [48].

3.5. Systèmes à événements discrets/Discrete event systems

Des systèmes dynamiques max-plus linéaires, de type (2), interviennent aussi, avec une interprétation toute différente, dans la modélisation des systèmes à événements discrets. Dans ce contexte, on associe à chaque tâche répétitive, i, une fonction *compteur*, $v_i : \mathbb{R} \to \mathbb{N}$, telle que $v_i(t)$ compte le nombre cumulé d'occurrences de la tâche i jusqu'à l'instant t. Par exemple, dans un système de production, $v_i(t)$ compte le nombre de pièces d'un certain type produites jusqu'à l'instant t. Dans le cas le plus simple, qui dans le langage des réseaux de Petri, correspond à la sous-classe très étudiée des graphes d'événements temporisés [82], on obtient des équations min-plus linéaires analogues à (2). Cette observation, ou plutôt, l'observation duale faisant intervenir des fonctions dateurs, a été le point de départ [86] de l'approche max-plus des systèmes à événements discrets [6], qui fournit un analogue max-plus de la théorie des systèmes linéaires classiques, incluant les notions de représentation d'état, de stabilité, de séries de transfert, etc. En particulier, les valeurs propres fournissent des mesures de performance telles que le taux de production. Des généralisations non-linéaires, telles que les systèmes dynamiques min-max [150], [123], ont aussi été étudiées. Les systèmes dynamiques max-plus linéaires aléatoires sont particulièrement utiles dans la modélisation des réseaux [67]. Les modèles d'automates à multiplicités max-plus [109], incluant certains versions temporisées des modèles de traces ou de tas de pièces [113], permettent de représenter des phénomènes de concurrence ou de partage de ressources. Les automates à multiplicités max-plus on été très étudiés par ailleurs en informatique théorique [161], [124], [137], [162], [131], [151]. Ils fournissent des modèles particulièrement adaptés à l'analyse de problèmes d'ordonnancement [136].

English version

Dynamical systems of type (2) also arise, with a different interpretation, in the modelling of discrete event systems. In this context, one associates to every repetitive task, i, a counter function, $v_i : \mathbb{R} \to \mathbb{N}$, such that $v_i(t)$ gives the total number of occurrences of task i up to time t. For instance, in a manufacturing system, $v_i(t)$ will count the number of parts of a given type produced up to time t. In the simplest case, which, in the vocabulary of Petri nets, corresponds to the much studied subclass of timed event graphs [82], we get min-plus linear equations similar to (2). This observation, or rather, the dual observation concerning dater functions, was the starting point [86] of the max-plus approach of discrete event systems [6], which provides some analogue of the classical linear control theory, including notions of state space representations, stability, transfer series, etc. In particular, eigenvalues yield performance measures like the throughput. Nonlinear generalisations, like min-max dynamical systems [150], [123], have been particularly studied. Random max-plus linear dynamical systems are particularly useful in the modelling of networks [67]. Max-plus automata models [109], which include some timed version of trace or heaps of pieces models [113], allow to represent phenomena of concurrency or resource sharing. Note that max-plus automata have been much studied in theoretical computer science [161], [124], [137], [162], [131], [151]. Such automata models are particularly adapted to the analysis of scheduling problems [136].

3.6. Algèbre linéaire max-plus/Basic max-plus algebra

Une bonne partie des résultats de l'algèbre max-plus concerne l'étude des systèmes d'équations linéaires. On peut distinguer trois familles d'équations, qui sont traitées par des techniques différentes : 1) Nous avons déjà évoqué dans les sections 3.2 et 3.3 le problème spectral max-plus $Ax = \lambda x$ et ses généralisations. Celui-ci apparaît en contrôle optimal déterministe et dans l'analyse des systèmes à événements discrets. 2) Le problème $Ax = b$ intervient en commande juste-à-temps (dans ce contexte, le vecteur x représente les dates de démarrage des tâches initiales, b représente certaines dates limites, et on se contente souvent de l'inégalité $Ax \leq b$). Le problème $Ax = b$ est intimement lié au problème d'affectation optimale, et plus généralement au problème de transport optimal. Il se traite via la théorie des correspondances de Galois abstraites, ou théorie de la résiduation [103], [74], [166], [169],[6]. Les versions dimension infinie du problème $Ax = b$ sont reliées aux questions d'analyse convexe abstraite [163], [158], [46] et de dualité non convexe. 3) Le problème linéaire général $Ax = Bx$ conduit à des développements combinatoires intéressants (polyèdres max-plus, déterminants max-plus, symétrisation [122], [152],[6]). Le sujet fait l'objet d'un intérêt récemment renouvelé [97].

English version

An important class of results in max-plus algebra concerns the study of max-plus linear equations. One can distinguish three families of equations, which are handled using different techniques: 1) We already mentioned in Sections 3.2 and 3.3 the max-plus spectral problem $Ax = \lambda x$ and its generalisations, which appears in deterministic optimal control and in performance analysis of discrete event systems. 2) The $Ax = b$ problem arises naturally in just in time problems (in this context, the vector x represents the starting times of initial tasks, b represents some deadlines, and one is often content with the inequality $Ax \leq b$). The $Ax = b$ problem is intimately related with optimal assignment, and more generally, with optimal transportation problems. Its theory relies on abstract Galois correspondences, or residuation theory [103], [74], [166], [169],[6]. Infinite dimensional versions of the $Ax = b$ problem are related to questions of abstract convex analysis [163], [158], [46] and nonconvex duality. 3) The general linear system $Ax = Bx$ leads to interesting combinatorial developments (max-plus polyhedra, determinants, symmetrisation [122], [152],[6]). The subject has attracted recently a new attention [97].

3.7. Algèbre max-plus et asymptotiques/Using max-plus algebra in asymptotic analysis

Le rôle de l'algèbre min-plus ou max-plus dans les problèmes asymptotiques est évident si l'on écrit

$$e^{-a/\epsilon} + e^{-b/\epsilon} \asymp e^{-\min(a,b)/\epsilon} \ , \qquad e^{-a/\epsilon} \times e^{-b/\epsilon} = e^{-(a+b)/\epsilon} \ , \tag{7}$$

lorsque $\epsilon \to 0^+$. Formellement, l'algèbre min-plus peut être vue comme la limite d'une déformation de l'algèbre classique, en introduisant le semi-anneau $\mathbb{R}_\epsilon$, qui est l'ensemble $\mathbb{R} \cup \{+\infty\}$, muni de l'addition $(a, b) \mapsto -\epsilon \log (e^{-a/\epsilon} + e^{-b/\epsilon})$ et de la multiplication $(a, b) \mapsto a + b$. Pour tout $\epsilon > 0$, $\mathbb{R}_\epsilon$ est isomorphe au semi-corps usuel des réels positifs, $(\mathbb{R}_+, +, \times)$, mais pour $\epsilon = 0^+$, $\mathbb{R}_\epsilon$ n'est autre que le semi-anneau min-plus. Cette idée a été introduite par Maslov [143], motivé par l'étude des asymptotiques de type WKB d'équations de Schrödinger. Ce point de vue permet d'utiliser des résultats algébriques pour résoudre des problèmes d'asymptotiques, puisque les équations limites ont souvent un caractère min-plus linéaire.

Cette déformation apparaît classiquement en théorie des grandes déviations à la loi des grands nombres : dans ce contexte, les objets limites sont des mesures idempotentes au sens de Maslov. Voir [1], [153], [47], pour les relations entre l'algèbre max-plus et les grandes déviations, voir aussi [42], [41], [40] pour des applications de ces idées aux perturbations singulières de valeurs propres. La même déformation est à l'origine de nombreux travaux actuels en géométrie tropicale, à la suite de Viro [165].

English version

The role of min-plus algebra in asymptotic problems becomes obvious when writing Equations (7) when $\epsilon \to 0^+$. Formally, min-plus algebra may be thought of as the limit of a deformation of classical algebra, by introducing the semi-field $\mathbb{R}_\epsilon$, which is the set $\mathbb{R} \cup \{+\infty\}$, equipped with the addition $(a, b) \mapsto -\epsilon \log (e^{-a/\epsilon} + e^{-b/\epsilon})$ and the multiplication $(a, b) \mapsto a + b$. For all $\epsilon > 0$, $\mathbb{R}_\epsilon$ is isomorphic to the semi-field of usual real positive numbers, $(\mathbb{R}_+, +, \times)$, but for $\epsilon = 0^+$, $\mathbb{R}_\epsilon$ coincides with the min-plus semiring. This idea was introduced by Maslov [143], motivated by the study of WKB-type asymptotics of Schrödinger equations. This point of view allows one to use algebraic results in asymptotics problems, since the limit equations have often some kind of min-plus linear structure.

This deformation appears classically in large deviation theory: in this context, the limiting objects are idempotent measures, in the sense of Maslov. See [1], [153], [47] for the relation between max-plus algebra and large deviations. See also [42], [41], [40] for the application of such ideas to singular perturbation problems for matrix eigenvalues. The same deformation is at the origin of many current works in tropical geometry, in the line initiated by Viro [165].

POEMS Project-Team

3. Research Program

3.1. General description

Our activity relies on the existence of boundary value problems established by physicists to model the propagation of waves in various situations. The basic ingredient is a linear partial differential equation of the hyperbolic type, whose prototype is the wave equation (or the Helmholtz equation if time-periodic solutions are considered). Nowadays, the numerical techniques for solving the basic academic problems are well mastered. However, the solution of complex wave propagation problems close to real applications still raises (essentially open) problems which constitute a real challenge for applied mathematicians. In particular, several difficulties arise when extending the results and the methods from the scalar wave equation to vectorial problems modeling wave propagation in electromagnetism or elastodynamics.

A large part of research in mathematics, when applied to wave propagation problems, is oriented towards the following goals:

- The design of new numerical methods, increasingly accurate and efficient.
- The development of artificial transparent boundary conditions for handling unbounded propagation domains.
- The treatment of more and more complex configurations (non local models, non linear models, coupled systems, periodic media).
- The study of specific phenomena such as guided waves and resonances, which raise mathematical questions of spectral theory.
- The development of approximate models via asymptotic analysis with multiple scales (thin layers, boundary layers effects, small homogeneities, homogenization, ...).
- The development and the analysis of algorithms for inverse problems (in particular for inverse scattering problems) and imaging techniques, using data from wave phenomena.

3.2. Wave propagation in non classical media

Extraordinary phenomena regarding the propagation of electromagnetic or acoustic waves appear in materials which have non classical properties: materials with a complex periodic microstructure that behave as materials with negative physical parameters, metals with a negative dielectric permittivity at optical frequencies, magnetized plasmas endowed with a strongly anisotropic and sign-indefinite permittivity tensor. These non classical materials raise original questions from theoretical and numerical points of view.

The objective is to study the well-posedness in this unusual context where physical parameters are sign-changing. New functional frameworks must be introduced, due, for instance, to hypersingularities of the electromagnetic field which appear at corners of metamaterials. This has of course numerical counterparts. In particular, classical Perfectly Matched Layers are unstable in these dispersive media, and new approaches must be developed.

Two ANR projects (METAMATH and CHROME) are related to this activity.

3.3. Wave propagation in heterogeneous media

Our objective is to develop efficient numerical approaches for the propagation of waves in heterogeneous media.

We aim on one hand to improve homogenized modeling of periodic media, by deriving enriched boundary conditions (or transmission conditions if the periodic structure is embedded in a homogeneous matrix) which take into account the boundary layer phenomena.

On the other hand, we like to develop multi-scale numerical methods when the assumption of periodicity on the spatial distribution of the heterogeneities is relaxed, or even completely lost. The general idea consists in a coupling between a macroscopic solver, based on a coarse mesh, with some microscopic representation of the field. This latter can be obtained by a numerical microscopic solver or by an analytical asymptotic expansion. This leads to two very different approaches which may be relevant for very different applications.

3.4. Spectral theory and modal approaches for waveguides

The study of waveguides is a longstanding and major topic of the team. Concerning the selfadjoint spectral theory for open waveguides, we turned recently to the very important case of periodic media. One objective is to design periodic structures with localized perturbations to create gaps in the spectrum, containing isolating eigenvalues.

Then, we would like to go further in proving the absence of localized modes in non uniform open waveguides. An original approach has been successfully applied to the scalar problem of a 2D junction. The challenge now is to extend these ideas to other configurations: 3D junctions, bent waveguides, vectorial problems...

Besides, we will continue our activity on modal methods for closed waveguides. In particular, we aim at extending the enriched modal method to take into account curvature and rough boundaries.

Finally, we are developing asymptotic models for networks of thin waveguides which arise in several applications (electric networks, simulation of lung, nanophotonics...).

3.5. Inverse problems

Building on the strong expertise of POEMS in the mathematical modeling of waves, most of our contributions aim at improving inverse scattering methodologies.

We acquired some expertise on the so called Linear Sampling Method, from both the theoretical and the practical points of view. Besides, we are working on topological derivative methods, which exploit small-defect asymptotics of misfit functionals and can thus be viewed as an alternative sampling approach, which can take benefit of our expertise on asymptotic methods.

An originality of our activity is to consider inverse scattering in waveguides (the inverse scattering community generally considers only free-space configurations). This is motivated at the same time by specific issues concerning the ill-posedness of the identification process and by applications to non-destructive techniques, for waveguide configurations (cables, pipes, plates etc...).

Lastly, we continue our work on the so-called exterior approach for solving inverse obstacle problems, which associates quasi-reversibility and level set methods. The objective is now to extend it to evolution problems.

3.6. Integral equations

Our activity in this field aims at developing accurate and fast methods for 3D problems.

On one hand, we developed a systematic approach to the analytical evaluation of singular integrals, which arise in the computation of the matrices of integral equations when two elements of the mesh are either touching each other or geometrically close.

On the other hand, POEMS is developing Fast Boundary Element Methods for 3D acoustics or elastodynamics, with applications to soil-structure interaction, seismology or seismic imaging.

Finally, a posteriori error analysis methodologies and adaptivity for boundary integral equation formulations of acoustic, electromagnetic and elastic wave propagation is investigated in the framework of the ANR project RAFFINE.

3.7. Domain decomposition methods

This is a come back to a topic in which POEMS contributed in the 1990's. It is motivated by our collaborations with the CEA-CESTA and the CEA-LIST, for the solution of large problems in time-harmonic electromagnetism and elastodynamics.

We combine in an original manner classical ideas of Domain Decomposition Methods with the specific formulations that we use for wave problems in unbounded domains, taking benefit of the available analytical representations of the solution (integral representation, modal expansion etc...).

SELECT Project-Team

3. Research Program

3.1. General presentation

From applications we treat on a day-to-day basis, we have learned that some assumptions currently used in asymptotic theory for model selection are often irrelevant in practice. For instance, it is not realistic to assume that the target belongs to the family of models in competition. Moreover, in many situations, it is useful to make the size of the model depend on the sample size, which makes asymptotic analyses breakdown. An important aim of SELECT is to propose model selection criteria which take such practical constraints into account.

3.2. A nonasymptotic view of model selection

An important goal of SELECT is to build and analyze penalized log-likelihood model selection criteria that are efficient when the number of models in competition grows to infinity with the number of observations. Concentration inequalities are a key tool for this, and lead to data-driven penalty choice strategies. A major research direction for SELECT consists of deepening the analysis of data-driven penalties, both from the theoretical and practical points of view. There is no universal way of calibrating penalties, but there are several different general ideas that we aim to develop, including heuristics derived from Gaussian theory, special strategies for variable selection, and resampling methods.

3.3. Taking into account the modeling purpose in model selection

Choosing a model is not only difficult theoretically. From a practical point of view, it is important to design model selection criteria that accommodate situations in which the data probability distribution P is unknown, and which take the model user's purpose into account. Most standard model selection criteria assume that P belongs to one of a set of models, without considering the purpose of the model. By also considering the model user's purpose, we can avoid or overcome certain theoretical difficulties, and produce flexible model selection criteria with data-driven penalties. The latter is useful in supervised classification and hidden-structure models.

3.4. Bayesian model selection

The Bayesian approach to statistical problems is fundamentally probabilistic: a joint probability distribution is used to describe the relationships among all unknowns and the data. Inference is then based on the posterior distribution, i.e., the conditional probability distribution of the parameters given the observed data. Exploiting the internal consistency of the probability framework, the posterior distribution extracts relevant information in the data and provides a complete and coherent summary of post-data uncertainty. Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle.

TAO Project-Team

3. Research Program

3.1. The Four Pillars of TAO

This Section describes TAO main research directions at the crossroad of Machine Learning and Evolutionary Computation. Since 2008, TAO has been structured in several special interest groups (SIGs) to enable the agile investigation of long-term or emerging theoretical and applicative issues. The comparatively small size of TAO SIGs enables in-depth and lively discussions; the fact that all TAO members belong to several SIGs, on the basis of their personal interests, enforces the strong and informal collaboration of the groups, and the fast information dissemination.

The first two SIGs consolidate the key TAO scientific pillars, while the others evolve and adapt to new topics.

The **Stochastic Continuous Optimization** SIG (OPT-SIG) takes advantage of the fact that TAO is acknowledged the best French research group and one of the top international groups in evolutionary computation from a theoretical and algorithmic standpoint. A main priority on the OPT-SIG research agenda is to provide theoretical and algorithmic guarantees for the current world state-of-the-art continuous stochastic optimizer, CMA-ES, ranging from convergence analysis (Youhei Akimoto's post-docs) to a rigorous benchmarking methodology. Incidentally, this benchmark platform COCO has been acknowledged since 2009 as "the" international continuous optimization benchmark, and its extension is at the core of the ANR project NumBBO (started end 2012). Another priority is to address the current limitations of CMA-ES in terms of high-dimensional or expensive optimization and constraint handling (respectively Ouassim Ait El Hara's, Ilya Loshchilov's PhDs and Asma Atamna's).

The **Optimal Decision Making under Uncertainty** SIG (UCT-SIG) benefits from the MoGo expertise and the team previous activity reports) and its past and present world records in the domain of computer-Go, establishing the international visibility of TAO in sequential decision making. Since 2010, UCT-SIG resolutely moves to address the problems of **energy management** from a fundamental and applied perspective. On the one hand, energy management offers a host of challenging issues, ranging from long-horizon policy optimization to the combinatorial nature of the search space, from the modeling of prior knowledge to non-stationary environment to name a few. On the other hand, the energy management issue can hardly be tackled in a pure academic perspective: tight collaborations with industrial partners are needed to access the true operational constraints. Such international and national collaborations have been started by Olivier Teytaud during his three stays (1 year, 6 months, 6 months) in Taiwan, and witnessed by the FP7 STREP Citines, the ADEME Post contract, and the METIS I-lab with SME Artelys.

The **Data Science** SIG (DS-SIG) now includes the activities related to the CDS and ISN Lidex in Saclay. On the one hand, it replaces and extends the former *Distributed systems* SIG, that was devoted to the modeling and optimization of (large scale) distributed systems, and itself was extending the goals of the original *Autonomic Computing* SIG, initiated by Cécile Germain-Renaud and investigating the use of statistical Machine Learning for large scale computational architectures, from data acquisition (the Grid Observatory in the European Grid Initiative) to grid management and fault detection. But these activities have become more and more application-driven, from High Energy Physics for the highly distributed computation to the Social Sciences for the multi-agents approaches – hence the change of focus of this SIG. A major result of this theme has been the creation 2 years ago of the Paris-Saclay Center for Data Science, co-chaired by Balázs Kégl, and the organization of the Higgs-ML challenge (http://higgsml.lal.in2p3.fr/), most popular challenge ever on the Kaggle platform.

On the other hand, several activities around Digital Humanities involving Gregory Grefenstette, Cécile Germain-Renaud, Michèle Sebag and Philippe Caillou, have widely extended previous work around the modeling of multi-agent systems and the exploitation of simulation results in the SimTools RNSC network frame. Digital Humanities involves adding semantics to underspecified collections of societal information: in an historical perspective (as in the new TAO H2020 project, EHRI-II on holocaust archives, or in the Gregorius project on church history); or an economical and societal perspective (as in the Cartolabe and AMIQAP projects); or an individual perspective (as in the ongoing Personal Semantics project). The key challenge here is to use learning algorithms to find structure and extract knwoledge from poorly structured or unstructured information, and to provide intelligible results and/or means to interact with the user.

The **Designing Criteria** SIG (CRI-SIG) focuses on the design of learning and optimization criteria. It elaborates on the lessons learned from the former *Complex Systems* SIG, showing that the key issue in challenging applications often is to design the objective itself. Such targeted criteria are pervasive in the study and building of autonomous cognitive systems, ranging from intrinsic rewards in robotics to the notion of saliency in vision and image understanding, and that of automatic algorithm selection and parameterization. The desired criteria can also result from fundamental requirements, such as scale invariance in a statistical physics perspective, and guide the algorithmic design. Additionally, the criteria can also be domain-driven and reflect the expert priors concerning the structure of the sought solution (e.g., spatio-temporal consistency); the challenge is to formulate such criteria in a mixed non convex/non differentiable objective function, nevertheless amenable to tractable optimization.

The **Deep Learning and Information Theory** SIG (DEEP-SIG) involves Yann Ollivier, Guillaume Charpiat, Michèle Sebag. This SIG originated from some extensions of the wwork done in the *Distributed Systems* SIG that have been developped in the context of the TIMCO FUI project (started end 2012 and just ended); the challenge was not only to port ML algorithms on massively distributed architectures, but to see how these architectures can inspire new ML criteria and methodologies. The coincidence of this project with the arrival of Yann Ollivier in TAO gradualy lead this work toward Deep Networks. This year, in addition to studying various theoretical and practical aspects of deep learning, we provide information-theoretic perspectives on the design and optimization of deep learning models, such as using the Fisher information matrix to optimize the parameters, or using minimum description length criteria to choose the right model structure (topology of the neural graph, addition or removal of parameters...) and to provide regularization and model selection.

AMIB Project-Team

3. Research Program

3.1. RNA

At the secondary structure level, we contributed novel generic techniques applicable to dynamic programming and statistical sampling, and applied them to design novel efficient algorithms for probing the conformational space. Another originality of our approach is that we cover a wide range of scales for RNA structure representation. For each scale (atomic, sequence, secondary and tertiary structure...) cutting-edge algorithmic strategies and accurate and efficient tools have been developed or are under development. This offers a new view on the complexity of RNA structure and function that will certainly provide valuable insights for biological studies.

3.1.1. Dynamic programming and complexity

Participants: Yann Ponty, Antoine Soulé.

*Common activity with J. Waldispühl (McGill) and A. Denise (*LRI*).*

Ever since the seminal work of Zuker and Stiegler, the field of RNA bioinformatics has been characterized by a strong emphasis on the secondary structure. This discrete abstraction of the 3D conformation of RNA has paved the way for a development of quantitative approaches in RNA computational biology, revealing unexpected connections between combinatorics and molecular biology. Using our strong background in enumerative combinatorics, we propose generic and efficient algorithms, both for sampling and counting structures using dynamic programming. These general techniques have been applied to study the sequence-structure relationship [56], the correction of pyrosequencing errors [48], and the efficient detection of multi-stable RNAs (riboswitches) [50], [51].

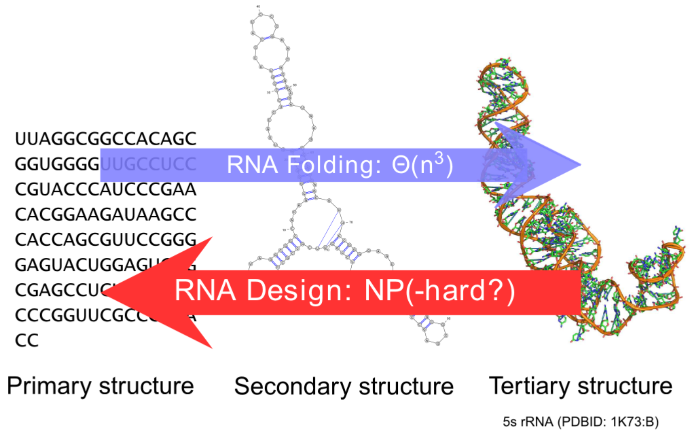


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alice Heliou, Vincent Le Gallic, Yann Ponty.

Joint project with A. Denise (sc Lri), S. Vialette (Marne-la-Vallée), J. Waldispühl (McGill) and Y. Zhang (Wuhan).

It is a natural pursue to build on our understanding of the secondary structure to construct artificial RNAs performing predetermined functions, ultimately targeting therapeutic and synthetic biology applications. Towards this goal, a key element is the design of RNA sequences that fold into a predetermined secondary structure, according to established energy models (inverse-folding problem). Quite surprisingly, and despite two decades of studies of the problem, the computational complexity of the inverse-folding problem is currently unknown.

Within our group, we offer a new methodology, based on weighted random generation [33] and multidimensional Boltzmann sampling, for this problem. Initially lifting the constraint of folding back into the target structure, we explored the random generation of sequences that are compatible with the target, using a probability distribution which favors exponentially sequences of high affinity towards the target. A simple posterior rejection step selects sequences that effectively fold back into the latter, resulting in a *global sampling* pipeline that showed comparable performances to its competitors based on local search [39].

3.1.3. Towards 3D modeling of large molecules

Participants: Yann Ponty, Afaf Saaidi, Mireille Regnier.

Joint projects with A. Denise (sc Lri), D. Barth (Versailles), J. Cohen (Paris-Sud), B. Sargueil (Paris V) and Jérome Waldispühl (McGill).

The modeling of large RNA 3D structures, that is predicting the three-dimensional structure of a given RNA sequence, relies on two complementary approaches. The approach by homology is used when the structure of a sequence homologous to the sequence of interest has already been resolved experimentally. The main problem then is to calculate an alignment between the known structure and the sequence. The ab initio approach is required when no homologous structure is known for the sequence of interest (or for some parts of it). We contribute methods inspired by both of these settings directions.

Modeling tasks can also be greatly helped by the availability of experimental data. However, high-resolution techniques such as crystallography or RMN, are notoriously costly in term of time and ressources, leading to the current gap between the amount of available sequences and structural data. As part of a colloboration with B. Sargueil's lab (Faculté de pharmacie, Paris V) funded by the Fondation pour la Recherche medical, we strive to propose a new paradigm for the analysis data produced using a new experimental technique, called SHAPE analysis (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension). This experimental setup produces an accessibility profile associated with the different positions of an RNA, the *shadow* of an RNA. As part of A. Saaidi's PhD, we currently design new algorithmic strategies to infer the secondary structure of RNA from multiple SHAPE experiments performed by experimentalists at Paris V. Those are obtained on mutants, and will be coupled with a fragment-based 3D modeling strategy developed by our partners at McGill.

3.2. Sequences

Participants: Mireille Régnier, Philippe Chassignet, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, Daria Iakovishina, Antoine Soulé.

String searching and pattern matching is a classical area in computer science, enhanced by potential applications to genomic sequences. In CPM/SPIRE community, a focus is given to general string algorithms and associated data structures with their theoretical complexity. Our group specialized in a formalization based on languages, weighted by a probabilistic model. Team members have a common expertise in enumeration and random generation of combinatorial sequences or structures, that are *admissible* according to some given constraints. A special attention is paid to the actual computability of formula or the efficiency of structures design, possibly to be reused in external software.

As a whole, motif detection in genomic sequences is a hot subject in computational biology that allows to address some key questions such as chromosome dynamics or annotation. Among specific motifs involved in molecular interactions, one may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). This area is being renewed by high throughput data and assembly issues. New constraints, such as energy conditions, or sequencing errors and amplification bias that are technology dependent, must be introduced in the models. A collaboration has beenestablished with LOB, at Ecole Polytechnique, who bought a sequencing machine, through the co-advised thesis of Alice Héliou. An other aim is to combine statistical sampling with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [44]. In general, in the future, our methods for sampling and sequence data analysis should be extended to take into account such constraints, that are continuously evolving.

3.2.1. Combinatorial Algorithms and motifs

Participants: Mireille Régnier, Philippe Chassignet, Alice Héliou, Daria Iakovishina.

Besides applications [5] of analytic combinatorics to computational biology problems, the team addressed general combinatorial problems on words and fundamental issues on languages and data structures.

Motif detection combines an algorithmic search of potential sites and a significance assessment. Assessment significance requires a quantitative criteriumsuch as the p-value.

In the recent years, a general scheme of derivation of analytic formula for the pvalue under different constraints (k-occurrence, first occurrence, overrepresentation in large sequences,...) has been provided. It relies on a representation of continuous sequences of overlapping words, currently named *clumps* or *clusters* in a graph [47]. Recursive equations to compute pvalues may be reduced to a traversal of that graph, leading to a linear algorithm. This improves over the space and time complexity of the generating function approach or previous probabilistic weighted automata.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly.

In [54], it is claimed that half of the genome consists of different types of repeats. One may cite microsatellites, DNA transposons, transposons, long terminal repeats (LTR), long interspersed nuclear elements (LINE), ribosomal DNA, short interspersed nuclear elements (SINE). Therefore, knowledge about the length of repeats is a key issue in several genomic problems, notably assembly or re-sequencing. Preliminary theoretical results are given in [37], and, recently, heuristics have been proposed and implemented [34], [49], [30]. A dual problem is the length of minimal absent words. Minimal absent words are words that do not occur but whose proper factors all occur in the sequence. Their computation is extremly related to finding maximal repeats (repeat that can not be extended on the right nor on the left). The comparison of the sets of minimal absent words provides a fast alternative for measuring approximation in sequence comparison [29], [32]. Recently, it was shown that considering the words which occur in one sequence but do no in another can be used to detect biologically significant events [52]. We have studied the computation of minimal absent words and we have provided new linear implementations [25],[21].

According to the current knowledge, cancer develops as a result of the mutational process of the genomic DNA. In addition to point mutations, cancer genomes often accumulate a significant number of chromosomal rearrangements also called structural variants (SVs). Identifying exact positions and types of these variants may lead to track cancer development or select the most appropriate treatment for the patient. Next Generation Sequencing opens the way to the study of structural variants in the genome, as recently described in [27]. This is the subject of an international collaboration with V. Makeev's lab (IOGENE, Moscow), MAGNOME project-team and V. Boeva (Curie Institute). One goal is to combine two detection techniquesbased either on paired-end mapping abnormalities or on variation of the depth of coverage. A second goal is to develop a model of errors, including a statistical model, that takes into account the quality of data from the different sequencing technologies, their volume and their specificities such as the GC-content or the mappability.

3.2.2. Random generation

Participant: Yann Ponty.

Analytical methods may fail when both sequential and structural constraints of sequences are to be modelled or, more generally, when molecular *structures* such as RNA structures have to be handled. The random generation of combinatorial objects is a natural, alternative, framework to assess the significance of observed phenomena. General and efficient techniques have been developed over the last decades to draw objects uniformly at random from an abstract specification. However, in the context of biological sequences and structures, the uniformity assumption becomes unrealistic, and one has to consider non-uniform distributions in order to derive relevant estimates. Typically, context-free grammars can handle certain kinds of long-range interactions such as base pairings in secondary RNA structures.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [35]: instead of formulating the problem as a classic optimization, this new approach uses statistical sampling within the space of solutions. Besides giving better, more robust, results, it allows for a fruitful adaptation of tools and algorithms derived in a purely combinatorial setting. Indeed, in a joint work with A. Denise (LRI), we have done significant and original progress in this area recently [46], [5], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [44].

3.3. 3D interaction and structure prediction

Participants: Julie Bernauer, Amélie Héliou.

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. This is specially challenging as structure flexibility is key and multi-scale modelling [26], [36] and efficient code are essential [40].

Our project covers various aspects of biological macromolecule structure and interaction modelling and analysis. First protein structure prediction is addressed through combinatorics. The dynamics of these types of structures is also studied using statistical and robotics inspired strategies. Both provide a good starting point to perform 3D interaction modelling, accurate structure and dynamics being essential.

Our group benefits from a good collaboration network, mainly at Stanford University (USA), HKUST (Hong-Kong) and McGill (Canada). The computational expertise in this field of computational structural biology is represented in a few large groups in the world (e.g. Pande lab at Stanford, Baker lab at U.Washington) that have both dry and wet labs. At Inria, our interest for structural biology is shared by the ABS and ORPAILLEUR project-teams. Our activities are however now more centered around protein-nucleic acid interactions, multi-scale analysis, robotics inspired strategies and machine learning than protein-protein interactions, algorithms and geometry. We also shared a common interest for large biomolecules and their dynamics with the NANO-D project team and their adaptative sampling strategy. As a whole, we contribute to the development of geometric and machine learning strategies for macromolecular docking.

Game theory was used by M. Boudard in her PhD thesis, defended in 2015, to predict the 3d structure of RNA. In her PhD thesis, co-advised by J. Cohen (LRI), A. Héliou is extending the approach to predict protein structures.

3.3.1. Statistical and robotics-inspired models for structure and dynamics

Participants: Julie Bernauer, Amélie Heliou.

Despite being able to correctly model small globular proteins, the computational structural biology community still craves for efficient force fields and scoring functions for prediction but also good sampling and dynamics strategies.

Our current and future efforts towards knowledge-based scoring function and ion location prediction have been described in 3.3.1 .

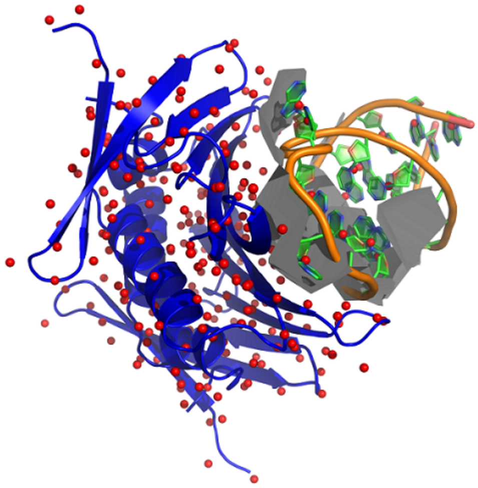


Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

Over the last two decades a strong connection between robotics and computational structural biology has emerged, in which internal coordinates of proteins are interpreted as a kinematic linkage with rotatable bonds as joints and corresponding groups of atoms as links [55], [31], [43], [42]. Initially, fragments in proteins limited to tens of residues were modeled as a kinematic linkage, but this approach has been extended to encompass (multi-domain) proteins [41]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [38], but it does not fully utilize the potential of a kinematic model. It breaks and recloses loops using six torsional degrees of freedom, which results in a finite number of solutions. The discrete nature of the solution set in the conformational space makes difficult an optimization of a target function with a gradient descent method. Our methods overcome this limitation by performing a conformational sampling and optimization in a co-dimension 6 subspace. Fragments remain closed, but these methods are limited to proteins. Our objective is to extend the approach proposed in [38], [55] to nucleic acids and protein/nucleic acid complexes with a view towards improving structure determination of nucleic acids and their complexes and in silico docking experiments of protein/RNA complexes. For that purpose, we have developed a generic strategy for differentiable statistical potentials [1], [53] that can be directly integrated in the procedure.

Results from in silico docking experiments will also directly benefit structure determination of complexes which, in turn, will provide structural insights in nucleic acid and protein/nucleic acid complexes. From the small proof-of-concept single chain protein implementation of the KGS strategy, we have developed a robust preliminary implementation that can handle RNA and will be further developed to account for multi-chain , with an extensive computational and biological validation.

GALEN Project-Team

3. Research Program

3.1. Shape, Grouping and Recognition

A general framework for the fundamental problems of image segmentation, object recognition and scene analysis is the interpretation of an image in terms of a set of symbols and relations among them. Abstractly stated, image interpretation amounts to mapping an observed image, X to a set of symbols Y. Of particular interest are the symbols Y^* that *optimally explain the underlying image*, as measured by a scoring function s that aims at distinguishing correct (consistent with human labellings) from incorrect interpretations:

$$Y^* = \mathrm{argmax}_Y s(X, Y) \tag{8}$$

Applying this framework requires (a) identifying which symbols and relations to use (b) learning a scoring function s from training data and (c) optimizing over Y in Eq.1 .

One of the main themes of our work is the development of methods that jointly address (a,b,c) in a shape-grouping framework in order to reliably extract, describe, model and detect shape information from natural and medical images. A principal motivation for using a shape-based framework is the understanding that shape- and more generally, grouping- based representations can go all the way from image features to objects. Regarding aspect (a), image representation, we cater for the extraction of image features that respect the shape properties of image structures. Such features are typically constructed to be purely geometric (e.g. boundaries, symmetry axes, image segments), or appearance-based, such as image descriptors. The use of machine learning has been shown to facilitate the robust and efficient extraction of such features, while the grouping of local evidence is known to be necessary to disambiguate the potentially noisy local measurements. In our research we have worked on improving feature extraction, proposing novel blends of invariant geometric- and appearance- based features, as well as grouping algorithms that allow for the efficient construction of optimal assemblies of local features.

Regarding aspect (b) we have worked on learning scoring functions for detection with deformable models that can exploit the developed low-level representations, while also being amenable to efficient optimization. Our works in this direction build on the graph-based framework to construct models that reflect the shape properties of the structure being modeled. We have used discriminative learning to exploit boundary- and symmetry-based representations for the construction of hierarchical models for shape detection, while for medical images we have developed methods for the end-to-end discriminative training of deformable contour models that combine low-level descriptors with contour-based organ boundary representations.

Regarding aspect (c) we have developed algorithms which implement top-down/bottom-up computation both in deterministic and stochastic optimization. The main idea is that ‘bottom-up’, image-based guidance is necessary for efficient detection, while ‘top-down’, object-based knowledge can disambiguate and help reliably interpret a given image; a combination of both modes of operation is necessary to combine accuracy with efficiency. In particular we have developed novel techniques for object detection that employ combinatorial optimization tools (A* and Branch-and-Bound) to tame the combinatorial complexity, achieving a best-case performance that is logarithmic in the number of pixels.

In the long run we aim at scaling up shape-based methods to 3D detection and pose estimation and large-scale object detection. One aspect which seems central to this is the development of appropriate mid-level representations. This is a problem that has received increased interest lately in the 2D case and is relatively mature, but in 3D it has been pursued primarily through ad-hoc schemes. We anticipate that questions pertaining to part sharing in 3D will be addressed most successfully by relying on explicit 3D representations. On the one hand depth sensors, such as Microsoft's Kinect, are now cheap enough to bring surface modeling and matching into the mainstream of computer vision - so these advances may be directly exploitable at test time for detection. On the other hand, even if we do not use depth information at test time, having 3D information can simplify the modeling task during training. In on-going work with collaborators we have started exploring combinations of such aspects, namely (i) the use of surface analysis tools to match surfaces from depth sensors (ii) using branch-and-bound for efficient inference in 3D space and (iii) groupwise-registration to build statistical 3D surface models. In the coming years we intend to pursue a tighter integration of these different directions for scalable 3D object recognition.

3.2. Machine Learning & Structured Prediction

The foundation of statistical inference is to learn a function that minimizes the expected loss of a prediction with respect to some unknown distribution

$$\mathcal{R}(f) = \int \ell(f, x, y) dP(x, y), \tag{9}$$

where $\ell(f, x, y)$ is a problem specific loss function that encodes a penalty for predicting $f(x)$ when the correct prediction is y. In our case, we consider x to be a medical image, and y to be some prediction, e.g. the segmentation of a tumor, or a kinematic model of the skeleton. The loss function, ℓ, is informed by the costs associated with making a specific misprediction. As a concrete example, if the true spatial extent of a tumor is encoded in y, $f(x)$ may make mistakes in classifying healthy tissue as a tumor, and mistakes in classifying diseased tissue as healthy. The loss function should encode the potential physiological damage resulting from erroneously targeting healthy tissue for irradiation, as well as the risk from missing a portion of the tumor.

A key problem is that the distribution P is unknown, and any algorithm that is to estimate f from labeled training examples must additionally make an implicit estimate of P. A central technology of empirical inference is to approximate $\mathcal{R}(f)$ with the empirical risk,

$$\mathcal{R}(f) \approx \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, x_i, y_i), \tag{10}$$

which makes an implicit assumption that the training samples (x_i, y_i) are drawn i.i.d. from P. Direct minimization of $\widehat{\mathcal{R}}(f)$ leads to overfitting when the function class $f \in \mathcal{F}$ is too rich, and regularization is required:

$$\min_{f \in \mathcal{F}} \lambda \Omega(\|f\|) + \widehat{\mathcal{R}}(f), \tag{11}$$

where Ω is a monotonically increasing function that penalizes complex functions.

Equation Eq. 4 is very well studied in classical statistics for the case that the output, $y \in \mathcal{Y}$, is a binary or scalar prediction, but this is not the case in most medical imaging prediction tasks of interest. Instead, complex interdependencies in the output space leads to difficulties in modeling inference as a binary prediction problem. One may attempt to model e.g. tumor segmentation as a series of binary predictions at each voxel in a medical image, but this violates the i.i.d. sampling assumption implicit in Equation Eq. 3 . Furthermore, we typically gain performance by appropriately modeling the inter-relationships between voxel predictions, e.g. by incorporating pairwise and higher order potentials that encode prior knowledge about the problem domain. It is in this context that we develop statistical methods appropriate to structured prediction in the medical imaging setting.

3.3. Self-Paced Learning with Missing Information

Many tasks in artificial intelligence are solved by building a model whose parameters encode the prior domain knowledge and the likelihood of the observed data. In order to use such models in practice, we need to estimate its parameters automatically using training data. The most prevalent paradigm of parameter estimation is supervised learning, which requires the collection of the inputs x_i and the desired outputs y_i. However, such an approach has two main disadvantages. First, obtaining the ground-truth annotation of high-level applications, such as a tight bounding box around all the objects present in an image, is often expensive. This prohibits the use of a large training dataset, which is essential for learning the existing complex models. Second, in many applications, particularly in the field of medical image analysis, obtaining the ground-truth annotation may not be feasible. For example, even the experts may disagree on the correct segmentation of a microscopical image due to the similarities between the appearance of the foreground and background.

In order to address the deficiencies of supervised learning, researchers have started to focus on the problem of parameter estimation with data that contains hidden variables. The hidden variables model the missing information in the annotations. Obtaining such data is practically more feasible: image-level labels ('contains car','does not contain person') instead of tight bounding boxes; partial segmentation of medical images. Formally, the parameters $\mathbf{w}$ of the model are learned by minimizing the following objective:

$$\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) + \sum_{i=1}^{n} \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{12}$$

Here, $\mathcal{W}$ represents the space of all parameters, n is the number of training samples, $R(\cdot)$ is a regularization function, and $\Delta(\cdot)$ is a measure of the difference between the ground-truth output y_i and the predicted output and hidden variable pair $(y_i(\mathbf{w}), h_i(\mathbf{w}))$.

Previous attempts at minimizing the above objective function treat all the training samples equally. This is in stark contrast to how a child learns: first focus on easy samples ('learn to add two natural numbers') before moving on to more complex samples ('learn to add two complex numbers'). In our work, we capture this intuition using a novel, iterative algorithm called self-paced learning (SPL). At an iteration t, SPL minimizes the following objective function:

$$\min_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in\{0,1\}^n} R(\mathbf{w}) + \sum_{i=1}^{n} v_i \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})) - \mu_t \sum_{i=1}^{n} v_i. \tag{13}$$

Here, samples with $v_i = 0$ are discarded during the iteration t, since the corresponding loss is multiplied by 0. The term μ_t is a threshold that governs how many samples are discarded. It is annealed at each iteration, allowing the learner to estimate the parameters using more and more samples, until all samples are used. Our results already demonstrate that SPL estimates accurate parameters for various applications such as image classification, discriminative motif finding, handwritten digit recognition and semantic segmentation. We will investigate the use of SPL to estimate the parameters of the models of medical imaging applications, such as segmentation and registration, that are being developed in the GALEN team. The ability to handle missing information is extremely important in this domain due to the similarities between foreground and background appearances (which results in ambiguities in annotations). We will also develop methods that are capable of minimizing more general loss functions that depend on the (unknown) value of the hidden variables, that is,

$$\min_{\mathbf{w}\in\mathcal{W},\theta\in\Theta} R(\mathbf{w}) + \sum_{i=1}^{n} \sum_{h_i\in\mathcal{H}} \Pr(h_i|x_i, y_i; \theta) \Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{14}$$

Here, θ is the parameter vector of the distribution of the hidden variables h_i given the input x_i and output y_i, and needs to be estimated together with the model parameters $\mathbf{w}$. The use of a more general loss function will allow us to better exploit the freely available data with missing information. For example, consider the case where y_i is a binary indicator for the presence of a type of cell in a microscopical image, and h_i is a tight bounding box around the cell. While the loss function $\Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to classify an image as containing a particular cell or not, the more general loss function $\Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to detect the cell as well (since h_i models its location)

3.4. Discrete Biomedical Image Perception

A wide variety of tasks in medical image analysis can be formulated as discrete labeling problems. In very simple terms, a discrete optimization problem can be stated as follows: we are given a discrete set of variables $\mathcal{V}$, all of which are vertices in a graph $\mathcal{G}$. The edges of this graph (denoted by $\mathcal{E}$) encode the variables' relationships. We are also given as input a discrete set of labels $\mathcal{L}$. We must then assign one label from $\mathcal{L}$ to each variable in $\mathcal{V}$. However, each time we choose to assign a label, say, x_{p_1} to a variable p_1, we are forced to pay a price according to the so-called *singleton* potential function $g_p(x_p)$, while each time we choose to assign a pair of labels, say, x_{p_1} and x_{p_2} to two interrelated variables p_1 and p_2 (two nodes that are connected by an edge in the graph $\mathcal{G}$), we are also forced to pay another price, which is now determined by the so called *pairwise* potential function $f_{p_1 p_2}(x_{p_1}, x_{p_2})$. Both the singleton and pairwise potential functions are problem specific and are thus assumed to be provided as input.

Our goal is then to choose a labeling which will allow us to pay the smallest total price. In other words, based on what we have mentioned above, we want to choose a labeling that minimizes the sum of all the MRF potentials, or equivalently the MRF energy. This amounts to solving the following optimization problem:

$$\arg\min_{\{x_p\}} \mathcal{P}(g,f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1,p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}). \tag{15}$$

The use of such a model can describe a number of challenging problems in medical image analysis. However these simplistic models can only account for simple interactions between variables, a rather constrained scenario for high-level medical imaging perception tasks. One can augment the expression power of this model through higher order interactions between variables, or a number of cliques $\{C_i, i \in [1,n] = \{\{p_{i^1}, \cdots, p_{i^{|C_i|}}\}\}$ of order $|C_i|$ that will augment the definition of $\mathcal{V}$ and will introduce hyper-vertices:

$$\arg\min_{\{x_p\}} \mathcal{P}(g,f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1,p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{E}} f_{p_1 \cdots p_n}(x_{p_{i^1}}, \cdots, p_{x_{i^{|C_i|}}}). \tag{16}$$

where $f_{p_1 \cdots p_n}$ is the price to pay for associating the labels $(x_{p_{i^1}}, \cdots, p_{x_{i^{|C_i|}}})$ to the nodes $(p_1 \cdots p_{i^{|C_i|}})$. Parameter inference, addressed by minimizing the problem above, is the most critical aspect in computational medicine and efficient optimization algorithms are to be evaluated both in terms of computational complexity as well as of inference performance. State of the art methods include deterministic and non-deterministic annealing, genetic algorithms, max-flow/min-cut techniques and relaxation. These methods offer certain strengths while exhibiting certain limitations, mostly related to the amount of interactions which can be tolerated among neighborhood nodes. In the area of medical imaging where domain knowledge is quite strong, one would expect that such interactions should be enforced at the largest scale possible.

M3DISIM Team

3. Research Program

3.1. Multi-scale modeling and coupling mechanisms for biomechanical systems, with mathematical and numerical analysis

Over the past decade, we have laid out the foundations of a multi-scale 3D model of the cardiac mechanical contraction responding to electrical activation. Several collaborations have been crucial in this enterprise, see below references. By integrating this formulation with adapted numerical methods, we are now able to represent the whole organ behavior in interaction with the blood during complete heart beats. This subject was our first achievement to combine a deep understanding of the underlying physics and physiology and our constant concern of proposing well-posed mathematical formulations and adequate numerical discretizations. In fact, we have shown that our model satisfies the essential thermo-mechanical laws, and in particular the energy balance, and proposed compatible numerical schemes that – in consequence – can be rigorously analyzed, see [5]. In the same spirit, we have recently formulated a poromechanical model adapted to the blood perfusion in the heart, hence precisely taking into account the large deformation of the mechanical medium, the fluid inertia and moving domain, and so that the energy balance between fluid and solid is fulfilled from the model construction to its discretization, see [6].

3.2. Inverse problems with actual data – Fundamental formulation, mathematical analysis and applications

A major challenge in the context of biomechanical modeling – and more generally in modeling for life sciences – lies in using the large amount of data available on the system to circumvent the lack of absolute modeling ground truth, since every system considered is in fact patient-specific, with possibly non-standard conditions associated with a disease. We have already developed original strategies for solving this particular type of inverse problems by adopting the observer stand-point. The idea we proposed consists in incorporating to the classical discretization of the mechanical system an estimator filter that can use the data to improve the quality of the global approximation, and concurrently identify some uncertain parameters possibly related to a diseased state of the patient, see [7], [8], [9]. Therefore, our strategy leads to a coupled model-data system solved similarly to a usual PDE-based model, with a computational cost directly comparable to classical Galerkin approximations. We have already worked on the formulation, the mathematical and numerical analysis of the resulting system – see [3] – and the demonstration of the capabilities of this approach in the context of identification of constitutive parameters for a heart model with real data, including medical imaging, see [1].

PARIETAL Project-Team

3. Research Program

3.1. Inverse problems in Neuroimaging

Many problems in neuroimaging can be framed as forward and inverse problems. For instance, the neuroimaging *inverse problem* consists in predicting individual information (behavior, phenotype) from neuroimaging data, while the *forward problem* consists in fitting neuroimaging data with high-dimensional (e.g. genetic) variables. Solving these problems entails the definition of two terms: a loss that quantifies the goodness of fit of the solution (does the model explain the data reasonably well ?), and a regularization schemes that represents a prior on the expected solution of the problem. In particular some priors enforce some properties of the solutions, such as sparsity, smoothness or being piece-wise constant.

Let us detail the model used in the inverse problem: Let $\mathbf{X}$ be a neuroimaging dataset as an (n_{subj}, n_{voxels}) matrix, where n_{subj} and n_{voxels} are the number of subjects under study, and the image size respectively, $\mathbf{Y}$ an array of values that represent characteristics of interest in the observed population, written as (n_{subj}, n_f) matrix, where n_f is the number of characteristics that are tested, and β an array of shape (n_{voxels}, n_f) that represents a set of pattern-specific maps. In the first place, we may consider the columns $\mathbf{Y}_1, .., \mathbf{Y}_{n_f}$ of Y independently, yielding n_f problems to be solved in parallel:

$$\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i, \forall i \in \{1, .., n_f\},$$

where the vector contains β_i is the i^{th} row of β. As the problem is clearly ill-posed, it is naturally handled in a regularized regression framework:

$$\widehat{\beta_i} = \text{argmin}_{\beta_i} \|\mathbf{Y}_i - \mathbf{X}\beta_i\|^2 + \Psi(\beta_i), \tag{17}$$

where Ψ is an adequate penalization used to regularize the solution:

$$\Psi(\beta; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 + \eta_1 \|\nabla\beta\|_1 + \eta_2 \|\nabla\beta\|_2 \tag{18}$$

with λ_1, λ_2, η_1, $\eta_2 \geq 0$ (this formulation particularly highlights the fact that convex regularizers are norms or quasi-norms). In general, only one or two of these constraints is considered (hence is enforced with a non-zero coefficient):

- When $\lambda_1 > 0$ only (LASSO), and to some extent, when $\lambda_1, \lambda_2 > 0$ only (elastic net), the optimal solution β is (possibly very) sparse, but may not exhibit a proper image structure; it does not fit well with the intuitive concept of a brain map.
- Total Variation regularization (see Fig. 1) is obtained for ($\eta_1 > 0$ only), and typically yields a piece-wise constant solution. It can be associated with Lasso to enforce both sparsity and sparse variations.
- Smooth lasso is obtained with ($\eta_2 > 0$ and $\lambda_1 > 0$ only), and yields smooth, compactly supported spatial basis functions.

The performance of the predictive model can simply be evaluated as the amount of variance in $\mathbf{Y}_i$ fitted by the model, for each $i \in \{1, .., n_f\}$. This can be computed through cross-validation, by *learning*$\widehat{\beta_i}$ on some part of the dataset, and then estimating $(Y_i - X\widehat{\beta_i})$ using the remainder of the dataset.

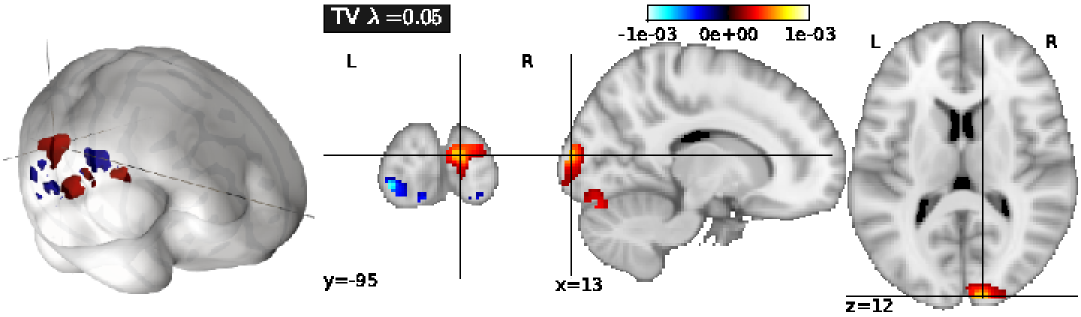


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here consists in predicting the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the observation of an image. Learning and test are performed across individuals. Unlike other approaches, Total Variation regularization yields a sparse and well-localized solution that enjoys particularly high accuracy.

This framework is easily extended by considering

- *Grouped penalization*, where the penalization explicitly includes a prior clustering of the features, i.e. voxel-related signals, into given groups. This is particularly important to include external anatomical priors on the relevant solution.
- *Combined penalizations*, i.e. a mixture of simple and group-wise penalizations, that allow some variability to fit the data in different populations of subjects, while keeping some common constraints.
- *Logistic regression*, where a logistic non-linearity is applied to the linear model so that it yields a probability of classification in a binary classification problem.
- *Robustness to between-subject variability* is an important question, as it makes little sense that a learned model depends dramatically on the particular observations used for learning. This is an important issue, as this kind of robustness is somewhat opposite to sparsity requirements.
- *Multi-task learning*: if several target variables are thought to be related, it might be useful to constrain the estimated parameter vector β to have a shared support across all these variables.
  For instance, when one of the variables $\mathbf{Y}_i$ is not well fitted by the model, the estimation of other variables $\mathbf{Y}_j, j \neq i$ may provide constraints on the support of β_i and thus, improve the prediction of $\mathbf{Y}_i$. Yet this does not impose constraints on the non-zero parameters of the parameters β_i.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{19}$$

then

$$\widehat{\beta} = \text{argmin}_{\beta=(\beta_i),i=1..n_f} \sum_{i=1}^{n_f} \|\mathbf{Y_i} - \mathbf{X}\beta_\mathbf{i}\|^2 + \lambda \sum_{j=1}^{n_{voxels}} \sqrt{\sum_{i=1}^{n_f} \beta_{\mathbf{i,j}}^\mathbf{2}} \tag{20}$$

3.2. Multivariate decompositions

Multivariate decompositions are an important tool to model complex data such as brain activation images: for instance, one might be interested in extracting an *atlas of brain regions* from a given dataset, such as regions depicting similar activities during a protocol, across multiple protocols, or even in the absence of protocol (during resting-state). These data can often be factorized into spatial-temporal components, and thus can be estimated through *regularized Principal Components Analysis* (PCA) algorithms, which share some common steps with regularized regression.

Let $\mathbf{X}$ be a neuroimaging dataset written as an (n_{subj}, n_{voxels}) matrix, after proper centering; the model reads

$$\mathbf{X} = \mathbf{A}\mathbf{D} + \epsilon, \tag{21}$$

where $\mathbf{D}$ represents a set of n_{comp} spatial maps, hence a matrix of shape (n_{comp}, n_{voxels}), and $\mathbf{A}$ the associated subject-wise loadings. While traditional PCA and independent components analysis are limited to reconstruct components $\mathbf{D}$ within the space spanned by the column of $\mathbf{X}$, it seems desirable to add some constraints on the rows of $\mathbf{D}$, that represent spatial maps, such as sparsity, and/or smoothness, as it makes the interpretation of these maps clearer in the context of neuroimaging.

This yields the following estimation problem:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{A}\mathbf{D}\|^2 + \Psi(\mathbf{D}) \text{ s.t. } \|\mathbf{A}_i\| = 1 \; \forall i \in \{1..n_f\}, \tag{22}$$

where $(\mathbf{A}_i), \; i \in \{1..n_f\}$ represents the columns of $\mathbf{A}$. Ψ can be chosen such as in Eq. (2) in order to enforce smoothness and/or sparsity constraints.

The problem is not jointly convex in all the variables but each penalization given in Eq (2) yields a convex problem on $\mathbf{D}$ for $\mathbf{A}$ fixed, and conversely. This readily suggests an alternate optimization scheme, where $\mathbf{D}$ and $\mathbf{A}$ are estimated in turn, until convergence to a local optimum of the criterion. As in PCA, the extracted components can be ranked according to the amount of fitted variance. Importantly, also, estimated PCA models can be interpreted as a probabilistic model of the data, assuming a high-dimensional Gaussian distribution (probabilistic PCA).

3.3. Covariance estimation

Another important estimation problem stems from the general issue of learning the relationship between sets of variables, in particular their covariance. Covariance learning is essential to model the dependence of these variables when they are used in a multivariate model, for instance to assess whether an observation is aberrant or not or in classification problems. Covariance learning is necessary to model latent interactions in high-dimensional observation spaces, e.g. when considering multiple contrasts or functional connectivity data.

The difficulties are two-fold: on the one hand, there is a shortage of data to learn a good covariance model from an individual subject, and on the other hand, subject-to-subject variability poses a serious challenge to the use of multi-subject data. While the covariance structure may vary from population to population, or depending on the input data (activation versus spontaneous activity), assuming some shared structure across problems, such as their sparsity pattern, is important in order to obtain correct estimates from noisy data. Some of the most important models are:

- **Sparse Gaussian graphical models**, as they express meaningful conditional independence relationships between regions, and do improve conditioning/avoid overfit.
- **Decomposable models**, as they enjoy good computational properties and enable intuitive interpretations of the network structure. Whether they can faithfully or not represent brain networks is an important question that needs to be addressed.
- **PCA-based regularization of covariance** which is powerful when modes of variation are more important than conditional independence relationships.

Adequate model selection procedures are necessary to achieve the right level of sparsity or regularization in covariance estimation; the natural evaluation metric here is the out-of-samples likelihood of the associated Gaussian model. Another essential remaining issue is to develop an adequate statistical framework to test differences between covariance models in different populations. To do so, we consider different means of parametrizing covariance distributions and how these parametrizations impact the test of statistical differences across individuals.

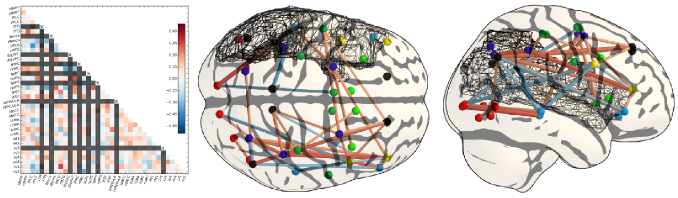


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical detection of the difference requires a sophisticated statistical framework for the comparison of graphical models.

POPIX Team

3. Research Program

3.1. Research Program

Mathematical models that characterize complex biological phenomena are complex numerical models which are defined by systems of ordinary differential equations when dealing with dynamical systems that evolve with respect to time, or by partial differential equations when there is a spatial component to the model. Also, it is sometimes useful to integrate a stochastic aspect into the dynamical systems in order to model stochastic intra-individual variability.

In order to use such methods, we are rapidly confronted with complex numerical difficulties, generally related to resolving the systems of differential equations. Furthermore, to be able to check the quality of a model, we require data. The statistical aspect of the model is thus critical in its way of taking into account different sources of variability and uncertainty, especially when data comes from several individuals and we are interested in characterizing the inter-subject variability. Here, the tool of reference is mixed-effects models.

Mixed-effects models are statistical models with both fixed effects and random effects, i.e., mixed effects. They are useful in many real-world situations, especially in the physical, biological and social sciences. In particular, they are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

POPIX develops new methods for estimation of complex mixed-effects models. Some of the extensions to these models that POPIX is actively researching include:

- models defined by a large system of differential equations
- models defined by a system of stochastic differential equations
- models defined by partial differential equations
- mixed hidden Markov models
- mixture models and model mixtures
- time-to-event models
- models including a large number of covariates

It is also important to clarify that POPIX is not meant to be a team of modelers; our main activity is not to develop models, but to develop tools for modelers. Indeed, we are of course led via our various collaborations to interact closely with modelers involved in model development, in particular in the case of our collaborations with modeling and simulation teams in the pharmaceutical industry. But POPIX is not in the business of building PKPD models per se.

Lastly, though pharmacometrics remains the main field of interest for the population approach, this approach is also appropriate to address other types of complex biological phenomena exhibiting inter-individual variability and necessitating therefore to be described by numerical and statistical models. We have already demonstrated the relevance of the developed approaches and tools in diverse other domains such as agronomy for characterizing corn production, and cellular biology for characterizing the cell cycle and the creation of free radicals in cells. Now we wish to push on to explore new areas of modeling such as for the respiratory system and blood flow. But again, it is not within the scope of the activities of POPIX to develop new models; instead, the goal is to demonstrate the relevance of the population approach in these areas.

INFINE Team

3. Research Program

3.1. Online Social Networks (OSN)

Large-scale online social networks such as Twitter or FaceBook provide a powerful means of selecting information. They rely on "social filtering", whereby pieces of information are collectively evaluated and sorted by users. This gives rise to information cascades when one item reaches a large population after spreading much like an epidemics from user to user in a viral manner. Nevertheless, such OSNs expose their users to a large amount of content of no interest to them, a sign of poor "precision" according to the terminology of information retrieval. At the same time, many more relevant content items never reach those users most interested in them. In other words, OSNs also suffer from poor "recall" performance.

This leads to a first challenge: *what determines the optimal trade-off between precision and recall in OSNs? And what mechanisms should be deployed in order to approach such an optimal trade-off?* We intend to study this question at a theoretical level, by elaborating models and analyses of social filtering, and to validate the resulting hypotheses and designs through experimentation and processing of data traces. More specifically, we envision to reach this general objective by solving the following problems.

3.1.1. Community Detection

Identification of implicit communities of like-minded users and contact recommendation for helping users "rewire" the information network for better performance. Potential schemes may include variants of spectral clustering and belief propagation-style message passing. Limitations / relative merits of candidate schemes, their robustness to noise in the input data, will be investigated.

3.1.2. Incentivization

Design of incentive mechanisms to limit the impact of users' selfishness on system behavior: efficiency should be maintained even when users are gaming the system to try and increase their estimated expertise. By offering rewards to users on the basis of their involvement in filtering and propagation of content, one might encourage them to adjust their action and contribute to increase the overall efficiency of the OSN as a content access platform.

One promising direction will be to leverage the general class of Vickrey-Clarke-Groves incentive-compatible mechanisms of economic theory to design so-called marginal utility reward mechanisms for OSN users.

3.1.3. Social Recommendation and Privacy

So far we have only alluded to the potential benefits of OSNs in terms of better information access. We now turn to the risks they create. Privacy breaches constitute the greatest of these risks: OSN users disclose a wealth of personal information and thereby expose themselves to discrimination by potential employers, insurers, lenders, government agencies...Such privacy concerns are not specific to OSNs: internauts' online activity is discretely tracked by companies such as Bluekai, and subsequently monetized to advertisers seeking better ad targeting. While disclosure of personal data creates a privacy risk, on the other hand it fuels personalized services and thereby potentially benefits everyone.

One line of research will be to focus on the specific application scenario of content categorization, and to characterize analytically the trade-off between user privacy protection (captured by differential privacy), accuracy of content categorization, and sample complexity (measured in number of probed users).

3.2. Traffic and resource management

Despite the massive increases in transmission capacity of the last few years, one has every reason to believe that networks will remain durably congested, driven among other factors by the steadily increasing demand for video content, the proliferation of smart devices (i.e., smartphones or laptops with mobile data cards), and the forecasted additional traffic due to machine-to-machine (M2M) communications. Despite this rapid traffic growth, there is still a rather limited understanding of the features protocols have to support, the characteristics of the traffic being carried and the context where it is generated. There is thus a strong need for smart protocols that transport requested information at the cheapest possible cost on the network as well as provide good quality of service to network subscribers. One particularly new aspect of up-and-coming networks is that networks are now used to not only (i) access information, but also (ii) distributively process information, en-route.

We intend to study these issues at the theoretical and protocol design levels, by elaborating models and analysis of content demands and/or mobility of network subscribers. The resulting hypothesis and designs will be validated through experimentation, simulation, or data trace processing. It is also worth mentioning the provided solutions may bring benefits to different entities in the network: to content owners (if applied at the core of Internet) or to subscribers or network operators (if applied at the edge of the Internet).

3.2.1. At the Internet Core

One important optimization variable consists in content replication: users can access the closest replica of the content they are interested in. Thus the memory resource can be used to create more replicas and reduce the usage of the bandwidth resource. Another interesting arbitrage between resources arises because content is no longer static but rather dynamic. Here are two simple examples: i) a video could be encoded at several resolutions. There is then a choice between pre-recording all possible resolutions, or alternatively synthesizing a lower-resolution version on the fly from a higher resolution version when a request arises. ii) A user requests the result of a calculation, say the average temperature in a building; this can either be kept in memory, or recomputed each time such a query arises. Optimizing the joint use of all three resources, namely bandwidth, memory, computation, is a complex task. Content Delivery Network companies such as Akamai or Limelight have worked on the memory/bandwidth trade-off for some years, but as we will explain more can be done on this. On the other hand optimizing the memory/computation trade-off has received far less attention. We aim to characterize the best possible content replication strategies by leveraging fine-grained prediction of i) users' future requests, and ii) wireless channels' future bandwidth fluctuations. In the past these two determining inputs have only been considered at a coarse-grained, aggregate level. It is important to assess how much bandwidth saving can be had by conducting finer-grained prediction. We are developing light-weight protocols for conducting these predictions and automatically instantiating the corresponding optimal replication policies. We are also investigating generic protocols for automatically trading replication for computation, focusing initially on the above video transcoding scenario.

3.2.2. At the Internet Edge

Cellular and wireless data networks are increasingly relied upon to provide users with Internet access on devices such as smartphones, laptops or tablets. In particular, the proliferation of handheld devices equipped with multiple advanced capabilities (e.g., significant CPU and memory capacities, cameras, voice to text, text to voice, GPS, sensors, wireless communication) has catalyzed a fundamental change in the way people are connected, communicate, generate and exchange data. In this evolving network environment, users' social relations, opportunistic resource availability, and proximity between users' devices are significantly shaping the use and design of future networking protocols.

One consequence of these changes is that mobile data traffic has recently experienced a staggering growth in volume: Cisco has recently foreseen that the mobile data traffic will increase 18-fold within 2016, in front of a mere 9-fold increase in connection speeds. Hence, one can observe today that the inherently centralized and terminal-centric communication paradigm of currently deployed cellular networks cannot cope with the increased traffic demand generated by smartphone users. This mismatch is likely to last because (1) forecasted mobile data traffic demand outgrows the capabilities of planned cellular technological advances such as 4G or LTE, and (2) there is strong skepticism about possible further improvements brought by 5G technology.

Congestion at the Internet's edge is thus here to stay. Solutions to this problem relates to: densify the infrastructure, opportunistically forward data among neighbors wireless devices, to offload data to alternate networks, or to bring content from the Internet closer to the subscribers. Our recent work on leveraging user mobility patterns, contact and inter-contact patterns, or content demand patterns constitute a starting point to these challenges. The projected increase of mobile data traffic demand pushes towards additional complementary offloading methods. Novel mechanisms are thus needed, which must fit both the new context that Internet users experience now, and their forecasted demands. In this realm, we will focus on new approaches leveraging ultra-distributed, user-centric approaches over IP.

3.3. Spontaneous Wireless Networks (SWN) and Internet of Things (IoT)

The unavailability of end-to-end connectivity in emergent wireless mobile networks is extremely disruptive for IP protocols. In fact, even in simpler cases of spontaneous wireless networks where end-to-end connectivity exists, such networks are still disruptive for the standard IP protocol stack, as many protocols rely on atomic link-local services (such as link-local multicast/broadcast), while these services are inherently unavailable in such networks due to their opportunistic, wireless multi hop nature. In this domain, we will aim to characterize the achievable performance in such IP-disruptive networks and to actively contribute to the design of new, deployable IP protocols that can tolerate these disruptions, while performing well enough compared to what is achievable and remaining interoperable with the rest of the Internet.

Spontaneous wireless networking is also a key aspect of the Internet of Things (IoT). The IoT is indeed expected to massively use this networking paradigm to gradually connect billions of new devices to the Internet, and drastically increase communication without human source or destination – to the point where the amount of such communications will dwarf communications involving humans. Large scale user environment automation require communication protocols optimized to efficiently leverage the heterogeneous and unreliable wireless vicinity (the scope of which may vary according to the application). In fact, extreme constraints in terms of cost, CPU, battery and memory capacities are typically experienced on a substantial fraction of IoT devices. We expect that such constraints will not vanish any time soon for two reasons. On one hand the progress made over the last decade concerning the cost/performance ratio for such small devices is quite disappointing. On the other hand, the ultimate goal of the IoT is ubiquitous Internet connectivity between devices as tiny as dust particles. These constraints actually require to redesign not only the network protocol stack running on these devices, but also the software platform powering these machines. In this context, we will aim at contributing to the design of novel network protocols and software platforms optimized to fit these constraints while remaining compatible with legacy Internet.

3.3.1. Design & Development of Open Experimental IoT Platforms

Based initially on "Demonstration abstract: Simply RIOT â Teaching and experimental research in the Internet of Things" Manufacturers announce on a regular basis the availability of novel tiny devices, most of them featuring network interfaces: the Internet of Things (IoT) is already here, from the hardware perspective, and it is expected in the near future that we will see a massive increase of the number of muti-purpose smart objects (from tiny sensors in industrial automation to devices like smart watches and tablets). Thus, one of the challenges is to be able to test architectures, protocols and applications, in realistic conditions and at large scale.

One necessity for research in this domain is to establish and improve IoT hardware platforms and testbeds, that integrate representative scenarios (such as Smart Energy, Home Automation etc.) and follow the evolution of technology, including radio technologies, and associated experimentation tools. For that, we plan to build upon the IoT-LAB federated testbeds, that we have participated in designing and deploying recently. We plan to further develop IoT-LAB with more heterogeneous, up-to-date IoT hardware and radios that will provide a usable and realistic experimentation environment. The goal is to provide a tool that enables testing an validation of upcoming software platforms and network stacks targeting concrete IoT deployments.

In parallel, on the software side, IoT hardware available so far made it uneasy for developers to build apps that run across heterogeneous hardware platforms. For instance Linux does not scale down to small, energy-constrained devices, while microcontroller-based OS alternatives were so far rudimentary and yield a steep learning curve and lengthy development life-cycles because they do not support standard programming and debugging tools. As a result, another necessity for research in this domain is to allow the emergence of it more powerful, unifying IOT software platforms, to bridge this gap. For that, we plan to build upon RIOT, a new open source software platform which provides a portable, Linux-like API for heterogeneous IoT hardware. We plan to continue to develop the systems and network stacks aspects of RIOT, within the open source developer community currently emerging around RIOT, which we co-founded together with Freie Universitaet Berlin. The key challenge is to improve usability and add functionalities, while maintaining architectural consistency and a small enough memory footprint. The goal is to provide an IoT software platform that can be used like Linux is used for less constrained machines, both (i) in the context of research and/or teaching, as well as (ii) in industrial contexts. Of course, we plan to use it ourselves for our own experimental research activities in the domain of IoT e.g., as an API to implement novel network protocols running on IoT hardware, to be tested and validated on IoT-LAB testbeds.

3.3.2. Design & Standardization of Architectures and Efficient Protocols for Internet of Things

As described before, and by definition, the Internet of Things will integrate not only a massive number of homogeneous devices (e.g., networks of wireless sensors), but also heterogeneous devices using various communication technologies. Most devices will be very constrained resources (memory resources, computational resources, energy). Communicating with (and amongst) such devices is a key challenge that we will focus on. The ability to communicate efficiently, to communicate reliably, or even just to be able to communicate at all, is non-trivial in many IoT scenarios: in this respect, we intend to develop innovative protocols, while following and contributing to standardization in this area. We will focus and base most of our work on standards developed in the context of the IETF, in working groups such as 6lo, CORE, LWIG etc., as well as IRTF research groups such as NWCRG on network coding and ICNRG on Information Centric Networking. We note however that this task goes far beyond protocol design: recently, radical rearchitecturing of the networks with new paradigms such as Information Centric Networking, ICN, (or even in wired networks, software-defined networks), have opened exciting new avenues. One of our direction of research will be to explore these content-centric approaches, and other novel architectures, in the context of IoT.

AVIZ Project-Team

3. Research Program

3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Visualization and Data Mining. Indirectly, it inherits from other established domains such as graphic design, Exploratory Data Analysis (EDA), statistics, Artificial Intelligence (AI), Human-Computer Interaction (HCI), and Psychology.

The use of graphic representation to understand abstract data is a goal Visual Analytics shares with Tukey's Exploratory Data Analysis (EDA) [59], graphic designers such as Bertin [48] and Tufte [58], and HCI researchers in the field of Information Visualization [47].

EDA is complementary to classical statistical analysis. Classical statistics starts from a *problem*, gathers *data*, designs a *model* and performs an *analysis* to reach a *conclusion* about whether the data follows the model. While EDA also starts with a problem and data, it is most useful *before* we have a model; rather, we perform visual analysis to discover what kind of model might apply to it. However, statistical validation is not always required with EDA; since often the results of visual analysis are sufficiently clear-cut that statistics are unnecessary.

Visual Analytics relies on a process similar to EDA, but expands its scope to include more sophisticated graphics and areas where considerable automated analysis is required before the visual analysis takes place. This richer data analysis has its roots in the domain of Data Mining, while the advanced graphics and interactive exploration techniques come from the scientific fields of Data Visualization and HCI, as well as the expertise of professions such as cartography and graphic designers who have long worked to create effective methods for graphically conveying information.

The books of the cartographer Bertin and the graphic designer Tufte are full of rules drawn from their experience about how the meaning of data can be best conveyed visually. Their purpose is to find effective visual representation that describe a data set but also (mainly for Bertin) to discover structure in the data by using the right mappings from abstract dimensions in the data to visual ones.

For the last 25 years, the field of Human-Computer Interaction (HCI) has also shown that interacting with visual representations of data in a tight perception-action loop improves the time and level of understanding of data sets. Information Visualization is the branch of HCI that has studied visual representations suitable to understanding and interaction methods suitable to navigating and drilling down on data. The scientific foundations of Information Visualization come from theories about perception, action and interaction.

Several theories of perception are related to information visualization such as the "Gestalt" principles, Gibson's theory of visual perception [52] and Triesman's "preattentive processing" theory [57]. We use them extensively but they only have a limited accuracy for predicting the effectiveness of novel visual representations in interactive settings.

Information Visualization emerged from HCI when researchers realized that interaction greatly enhanced the perception of visual representations.

To be effective, interaction should take place in an interactive loop faster than 100ms. For small data sets, it is not difficult to guarantee that analysis, visualization and interaction steps occur in this time, permitting smooth data analysis and navigation. For larger data sets, more computation should be performed to reduce the data size to a size that may be visualized effectively.

In 2002, we showed that the practical limit of InfoVis was on the order of 1 million items displayed on a screen [50]. Although screen technologies have improved rapidly since then, eventually we will be limited by the physiology of our vision system: about 20 millions receptor cells (rods and cones) on the retina. Another problem will be the limits of human visual attention, as suggested by our 2006 study on change blindness in large and multiple displays [49]. Therefore, visualization alone cannot let us understand very large data sets. Other techniques such as aggregation or sampling must be used to reduce the visual complexity of the data to the scale of human perception.

Abstracting data to reduce its size to what humans can understand is the goal of Data Mining research. It uses data analysis and machine learning techniques. The scientific foundations of these techniques revolve around the idea of finding a good model for the data. Unfortunately, the more sophisticated techniques for finding models are complex, and the algorithms can take a long time to run, making them unsuitable for an interactive environment. Furthermore, some models are too complex for humans to understand; so the results of data mining can be difficult or impossible to understand directly.

Unlike pure Data Mining systems, a Visual Analytics system provides analysis algorithms and processes compatible with human perception and understandable to human cognition. The analysis should provide understandable results quickly, even if they are not ideal. Instead of running to a predefined threshold, algorithms and programs should be designed to allow trading speed for quality and show the tradeoffs interactively. This is not a temporary requirement: it will be with us even when computers are much faster, because good quality algorithms are at least quadratic in time (e.g. hierarchical clustering methods). Visual Analytics systems need different algorithms for different phases of the work that can trade speed for quality in an understandable way.

Designing novel interaction and visualization techniques to explore huge data sets is an important goal and requires solving hard problems, but how can we assess whether or not our techniques and systems provide real improvements? Without this answer, we cannot know if we are heading in the right direction. This is why we have been actively involved in the design of evaluation methods for information visualization [56], [55], [53], [54], [51]. For more complex systems, other methods are required. For these we want to focus on longitudinal evaluation methods while still trying to improve controlled experiments.

3.2. Innovation

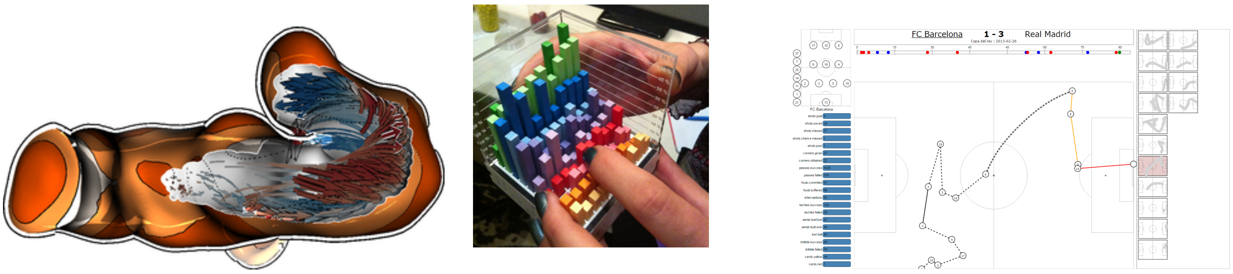


Figure 1. Example novel visualization techniques and tools developed by the team. Left: a non-photorealistic rendering technique that visualizes blood flow and vessel thickness. Middle:a physical visualization showing economic indicators for several countries, right: SoccerStories a tool for visualizing soccer games.

We design novel visualization and interaction techniques (see, for example, Figure 1). Many of these techniques are also evaluated throughout the course of their respective research projects. We cover application domains such as sports analysis, digital humanities, fluid simulations, and biology. A focus of Aviz' work is the improvement of graph visualization and interaction with graphs. We further develop individual techniques for the design of tabular visualizations and different types of data charts. Another focus is the use of animation as a transition aid between different views of the data. We are also interested in applying techniques from illustrative visualization to visual representations and applications in information visualization as well as scientific visualization.

3.3. Evaluation Methods

Evaluation methods are required to assess the effectiveness and usability of visualization and analysis methods. Aviz typically uses traditional HCI evaluation methods, either quantitative (measuring speed and errors) or qualitative (understanding users tasks and activities). Moreover, Aviz is also contributing to the improvement of evaluation methods by reporting on the best practices in the field, by co-organizing workshops (BELIV 2010, 2012, 2014, 2016) to exchange on novel evaluation methods, by improving our ways of reporting, interpreting and communicating statistical results, and by applying novel methodologies, for example to assess visualization literacy.

3.4. Software Infrastructures

We want to understand the requirements that software and hardware architectures should provide to support exploratory analysis of large amounts of data. So far, “big data” has been focusing on issues related to storage management and predictive analysis: applying a well-known set of operations on large amounts of data. Visual Analytics is about exploration of data, with sometimes little knowledge of its structure or properties. Therefore, interactive exploration and analysis is needed to build knowledge and apply appropriate analyses; this knowledge and appropriateness is supported by visualizations. However, applying analytical operations on large data implies long-lasting computations, incompatible with interactions, and generates large amounts of results, impossible to visualize directly without aggregation or sampling. Visual Analytics has started to tackle these problems for specific applications but not in a general manner, leading to fragmentation of results and difficulties to reuse techniques from one application to the other. We are interested in abstracting-out the issues and finding general architectural models, patterns, and frameworks to address the Visual Analytics challenge in more generic ways.

3.5. Emerging Technologies

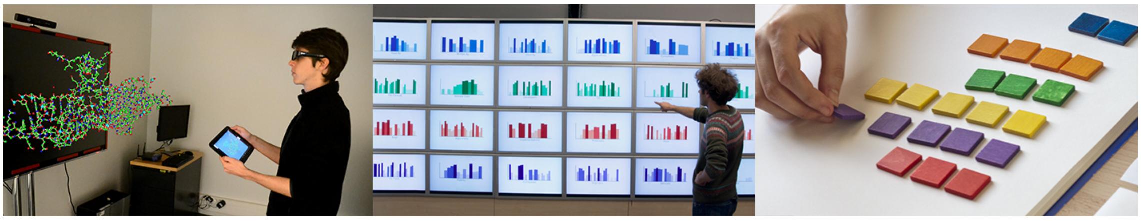


Figure 2. Example emerging technology solutions developed by the team for multi-display environments, wall displays, and token-based visualization.

We want to empower humans to make use of data using different types of display media and to enhance how they can understand and visually and interactively explore information. This includes novel display equipment and accompanying input techniques. The Aviz team specifically focuses on the exploration of the use of large displays in visualization contexts as well as emerging physical and tangible visualizations. In terms of interaction modalities our work focuses on using touch and tangible interaction. Aviz participates to the Digiscope project that funds 11 wall-size displays at multiple places in the Paris area (see http://www.

digiscope.fr), connected by telepresence equipment and a Fablab for creating devices. Aviz is in charge of creating and managing the Fablab, uses it to create physical visualizations, and is also using the local wall-size display (called WILD) to explore visualization on large screens. The team also investigates the perceptual, motor and cognitive implications of using such technologies for visualization.

3.6. Psychology

More cross-fertilization is needed between psychology and information visualization. The only key difference lies in their ultimate objective: understanding the human mind vs. helping to develop better tools. We focus on understanding and using findings from psychology to inform new tools for information visualization. In many cases, our work also extends previous work in psychology. Our approach to the psychology of information visualization is largely holistic and helps bridge gaps between perception, action and cognition in the context of information visualization. Our focus includes the perception of charts in general, perception in large display environments, collaboration, perception of animations, how action can support perception and cognition, and judgment under uncertainty.

DAHU Project-Team

3. Research Program

3.1. Research Program

Dahu aims at developing mechanisms for high-level specifications of systems built around DBMS, that are easy to understand while also facilitating verification of critical properties. This requires developing tools that are suitable for reasoning about systems that manipulate data. Some tools for specifying and reasoning about data have already been studied independently by the database community and by the verification community, with various motivations. However, this work is still in its infancy and needs to be further developed and unified.

Most current proposals for reasoning about DBMS over XML documents are based on tree automata, taking advantage of the tree structure of XML documents. For this reason, the Dahu team is studying a variety of tree automata. This ranges from restrictions of “classical” tree automata in order to understand their expressive power, to extensions of tree automata in order to understand how to incorporate the manipulation of data.

Moreover, Dahu is also interested in logical frameworks that explicitly refer to data. Such logical frameworks can be used as high level declarative languages for specifying integrity constraints, format change during data exchange, web service functionalities and so on. Moreover, the same logical frameworks can be used to express the critical properties we wish to verify.

In order to achieve its goals, Dahu brings together world-class expertise in both databases and verification.

EX-SITU Team

3. Research Program

3.1. Research Program

We characterize Extreme Situated Interaction as follows:

Extreme users. We study extreme users who make extreme demands on current technology. We know that human beings take advantage of the laws of physics to find creative new uses for physical objects. However, this level of adaptability is severely limited when manipulating digital objects. Even so, we find that creative professionals—artistists, designers and scientists—often adapt interactive technology in novel and unexpected ways and find creative solutions. By studying these users, we hope to not only address the specific problems they face, but also to identify the underlying principles that will help us to reinvent virtual tools. We seek to shift the paradigm of interactive software, to establish the laws of interaction that significantly empower users and allow them to control their digital environment.

Extreme situations. We develop extreme environments that push the limits of today's technology. We take as given that future developments will solve “practical" problems such as cost, reliability and performance and concentrate our efforts on interaction in and with such environments. This has been a successful strategy in the past: Personal computers only became prevalent after the invention of the desktop graphical user interface. Smartphones and tablets only became commercially successful after Apple cracked the problem of a usable touch-based interface for the iPhone and the iPad. Although wearable technologies, such as watches and glasses, are finally beginning to take off, we do not believe that they will create the major disruptions already caused by personal computers, smartphones and tablets. Instead, we believe that future disruptive technologies will include fully interactive paper and large interactive displays.

Our extensive experience with the Digiscope WILD and WILDER platforms places us in a unique position to understand the principles of distributed interaction that extreme environments call for. We expect to integrate, at a fundamental level, the collaborative capabilities that such environments afford. Indeed almost all of our activities in both the digital and the physical world take place within a complex web of human relationships. Current systems only support, at best, passive sharing of information, e.g., through the distribution of independent copies. Our goal is to support active collaboration, in which multiple users are actively engaged in the lifecycle of digital artifacts.

Extreme design. We explore novel approaches to the design of interactive systems, with particular emphasis on extreme users in extreme environments. Our goal is to empower creative professionals, allowing them to act as both designers and developers throughout the design process. Extreme design affects every stage, from requirements definition, to early prototyping and design exploration, to implmentation, to adaptation and appropriation by end users. We hope to push the limits of participatory design to actively support creativity at all stages of the design lifecycle.

Extreme design does not stop with purely digital artifacts. The advent of digital fabrication tools and FabLabs has significantly lowered the cost of making physical objects interactive. Creative professionals now create hybrid interactive objects that can be tuned to the user's needs. Integrating the design of physical objects into the software design process raises new challenges, with new methods and skills to support this form of extreme prototyping.

Our overall approach is to identify a small number of specific projects, organized around four themes: *Creativity, Augmentation, Collaboration* and *Infrastructure*. Specific projects may address multiple themes, and different members of the group work together to advance these different topics.

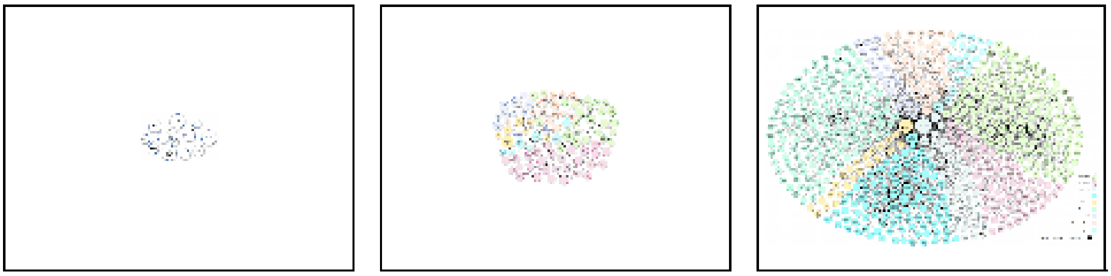
ILDA Team

3. Research Program

3.1. Introduction

Our ability to acquire or generate, store, process, interlink and query data has increased spectacularly over the last few years. The corresponding advances are commonly grouped under the umbrella of so called *Big Data*. Even if the latter has become a buzzword, these advances are real, and they are having a profound impact in domains as varied as scientific research, commerce, social media, industrial processes or e-government. Yet, looking ahead, emerging technologies related to what we now call the *Web of Data* (a.k.a the Semantic Web) have the potential to create an even larger revolution in data-driven activities, by making information accessible to machines as semistructured data [22] that eventually becomes actionable knowledge. Indeed, novel Web data models considerably ease the interlinking of semi-structured data originating from multiple independent sources. They make it possible to associate machine-processable semantics with the data. This in turn means that heterogeneous systems can exchange data, infer new data using reasoning engines, and that software agents can cross data sources, resolving ambiguities and conflicts between them [64]. Datasets are becoming very rich and very large. They are gradually being made even larger and more heterogeneous, but also much more useful, by interlinking them, as exemplified by the Linked Data initiative [41].

These advances raise research questions and technological challenges that span numerous fields of computer science research: databases, communication networks, security and trust, data mining, as well as human-computer interaction. Our research is based on the conviction that interactive systems play a central role in many data-driven activity domains. Indeed, no matter how elaborate the data acquisition, processing and storage pipelines are, data eventually get processed or consumed one way or another by users. The latter are faced with large, increasingly interlinked heterogeneous datasets (see, e.g., Figure 1) that are organized according to complex structures, resulting in overwhelming amounts of both raw data and structured information. Users thus require effective tools to make sense of their data and manipulate them.



2007: 12 linked datasets 2009: 93 linked datasets 2014: 570 linked datasets

Figure 1. Linking Open Data cloud diagram from 2007 to 2014 – http://lod-cloud.net

We approach this problem from the perspective of the Human-Computer Interaction (HCI) field of research, whose goal is to study how humans interact with computers and inspire novel hardware and software designs aimed at optimizing properties such as efficiency, ease of use and learnability, in single-user or cooperative work contexts. More formally, HCI is about designing systems that lower the barrier between users' cognitive model of what they want to accomplish, and computers' understanding of this model. HCI is about the design, implementation and evaluation of computing systems that humans interact with [46], [66]. It is a highly multidisciplinary field, with experts from computer science, cognitive psychology, design, engineering, ethnography, human factors and sociology.

In this broad context, ILDA aims at designing interactive systems that display [31], [53], [73] the data and let users interact with them, aiming to help users better *navigate* and *comprehend* large webs of data represented visually, as well as *relate* and *manipulate* them.

Our research agenda consists of the three complementary axes detailed in the following subsections. Designing systems that consider interaction in close conjunction with data semantics is pivotal to all three axes. Those semantics will help drive navigation in, and manipulation of, the data, so as to optimize the communication bandwidth between users and data.

3.2. Semantics-driven Data Manipulation

Participants: Emmanuel Pietriga, Caroline Appert, Hande Ozaygen, Mengying Du, Hugo Romat.

The Web of Data has been maturing for the last fifteen years and is starting to gain adoption across numerous application domains (Figure 1). Now that most foundational building blocks are in place, from knowledge representation, inference mechanisms and query languages [42], all the way up to the expression of data presentation knowledge [60] and to mechanisms like look-up services [72] or spreading activation [37], we need to pay significant attention to how human beings are going to interact with this new Web, if it is to *"reach its full potential"* [38].

Most efforts in terms of user interface design and development for the Web of data have essentially focused on tools for software developers or subject-matter experts who create ontologies and populate them [48], [36]. Tools more oriented towards end-users are starting to appear [28], [30], [43], [44], [47], [55], including the so-called *linked data browsers* [41]. However, those browsers are in most cases based on quite conventional point-and-click hypertext interfaces that present data to users in a very page-centric, web-of-documents manner that is ill-suited to navigating in, and manipulating, webs of data.

To be successful, interaction paradigms that let users navigate and manipulate data on the Web have to be tailored to the radically different way of browsing information enabled by it, where users directly interact with the data rather than with monolithic documents. The general research question addressed in this part of our research program is how to design novel interaction techniques that help users manipulate their data more efficiently. By data manipulation, we mean all low-level tasks related to manually creating new content, modifying and cleaning existing content, merging data from different sources, establishing connections between datasets, categorizing data, and eventually sharing the end results with other users; tasks that are currently considered quite tedious because of the sheer complexity of the concepts, data models and syntax, and the interplay between all of them.

Our approach is based on the conviction that there is a strong potential for cross-fertilization, as mentioned earlier: on the one hand, user interface design is essential to the management and understanding of webs of data; on the other hand, interlinked datasets enriched with even a small amount of semantics can help create more powerful user interfaces, that provide users with the right information at the right time.

We envision systems that focus on the data themselves, exploiting the underlying *semantics and structure* in the background rather than exposing them – which is what current user interfaces for the Web of Data often do. We envision interactive systems in which the semantics and structure are not exposed directly to users, but serve as input to the system to generate interactive representations that convey information relevant to the task at hand and best afford the possible manipulation actions.

3.3. Generalized Multi-scale Navigation

Participants: Olivier Chapuis, Emmanuel Pietriga, Caroline Appert, Anastasia Bezerianos, Olivier Gladin, María-Jesús Lobo, Arnaud Prouzeau.

The foundational question addressed here is what to display when, where and how, so as to provide effective support to users in their data understanding and manipulation tasks. ILDA targets contexts in which workers have to interact with complementary views on the same data, or with views on different-but-related datasets, possibly at different levels of abstraction. Being able to combine or switch between representations of the data at different levels of detail and merge data from multiple sources in a single representation is central to many scenarios. This is especially true in both of the application domains we consider: mission-critical systems (e.g., natural disaster crisis management) and the exploratory analysis of scientific data (e.g., correlate theories and heterogeneous observational data for an analysis of a given celestial body in Astrophysics).

A significant part of our research over the last ten years has focused on multi-scale interfaces. We designed and evaluated novel interaction techniques, but also worked actively on the development of open-source UI toolkits for multi-scale interfaces (see Section 6.2). These interfaces let users navigate large but relatively homogeneous datasets at different levels of detail, on both workstations [8], [25], [59], [58], [57], [26], [62], [24], [63] and wall-sized displays [5], [49], [61], [54], [27], [33], [32]. This part of the ILDA research program is about extending multi-scale navigation in two directions: 1. Enabling the representation of multiple, spatially-registered but widely varying, multi-scale data layers in Geographical Information Systems (GIS); 2. Generalizing the multi-scale navigation paradigm to interconnected, heterogeneous datasets as found on the Web of Data.

The first research problem is mainly investigated in collaboration with IGN in the context of ANR project MapMuxing (Section 8.2.1), which stands for *multi-dimensional map multiplexing*. Project MapMuxing aims at going beyond the traditional pan & zoom and overview+detail interface schemes, and at designing and evaluating novel cartographic visualizations that rely on high-quality generalization, *i.e.*, the simplification of geographic data to make it legible at a given map scale [69], [70], and symbol specification. Beyond project MapMuxing, we are also investigating multi-scale multiplexing techniques for geo-localized data in the specific context of ultra-high-resolution wall-sized displays, where the combination of a very high pixel density and large physical surface (Figure 2) enable us to explore designs that involve collaborative interaction and physical navigation in front of the workspace. This is work done in cooperation with team Massive Data at Inria Chile.

The second research problem is about the extension of multi-scale navigation to interconnected, heterogeneous datasets. Generalization has a rather straightforward definition in the specific domain of geographical information systems, where data items are geographical entities that naturally aggregate as scale increases. But it is unclear how generalization could work for representations of the more heterogeneous webs of data that we consider in the first axis of our research program. Those data form complex networks of resources with multiple and quite varied relationships between them, that cannot rely on a single, unified type of representation (a role played by maps in GIS applications).

Addressing the limits of current generalization processes is a longer-term, more exploratory endeavor. Here again, the machine-processable semantics and structure of the data give us an opportunity to rethink how users navigate interconnected heterogeneous datasets. Using these additional data, we investigate ways to generalize the multi-scale navigation paradigm to datasets whose layout and spatial relationships can be much richer and much more diverse than what can be encoded with static linear hierarchies as typically found today in interfaces for browsing maps or large imagery. Our goal is thus to design and develop highly dynamic and versatile multi-scale information spaces for heterogeneous data whose structure and semantics are not known in advance, but discovered incrementally.

3.4. Novel Forms of Input for Groups and Individuals

Participants: Caroline Appert, Anastasia Bezerianos, Olivier Chapuis, Emmanuel Pietriga, André Spritzer, Can Liu, Rafael Morales Gonzalez, Bruno Fruchard, Hae Jin Song.

Analyzing and manipulating large datasets can involve multiple users working together in a coordinated manner in multi-display environments: workstations, handheld devices, wall-sized displays [27]. Those users work towards a common goal, navigating and manipulating data displayed on various hardware surfaces in a coordinated manner. Group awareness [40], [21] is central in these situations, as users, who may or may not be co-located in the same room, can have an optimal individual behavior only if they have a clear picture of what their collaborators have done and are currently doing in the global context. We work on the design and implementation of interactive systems that improve group awareness in co-located situations [50], making individual users able to figure out what other users are doing without breaking the flow of their own actions.

In addition, users need a rich interaction vocabulary to handle large, structured datasets in a flexible and powerful way, regardless of the context of work. Input devices such as mice and trackpads provide a limited number of input actions, thus requiring users to switch between modes to perform different types of data manipulation and navigation actions. The action semantics of these input devices are also often too much dependent on the display output. For instance, a mouse movement and click can only be interpreted according to the graphical controller (widget) above which it is moved. We focus on designing powerful input techniques based upon technologies such as tactile surfaces (supported by UI toolkits developed in-house), 3D motion tracking systems, or custom-built controllers [52] *to complement (rather than replace) traditional input devices* such as keyboards, that remain the best method so far for text entry, and indirect input devices such as mice or trackpads for pixel-precise pointing actions.

The input vocabularies we investigate enable users to navigate and manipulate large and structured datasets in environments that involve multiple users and displays that vary in their size, position and orientation [27], [39], each having their own characteristics and affordances: wall displays [5], [74], workstations, tabletops [56], [35], tablets [7], [71], smartphones [10], [34], [67], [68], and combinations thereof [3], [9], [54], [27].

We aim at designing rich interaction vocabularies that go far beyond what current touch interfaces offer, which rarely exceeds five gestures such as simple slides and pinches. Designing larger gesture vocabularies requires identifying discriminating dimensions (e.g., the presence or absence of anchor points and the distinction between internal and external frames of reference [7]) in order to structure a space of gestures that interface designers can use as a dictionary for choosing a coherent set of controls. These dimensions should be few and simple, so as to provide users with gestures that are easy to memorize and execute. Beyond gesture complexity, the scalability of vocabularies also depends on our ability to design robust gesture recognizers that will allow users to fluidly chain simple gestures that make it possible to interlace navigation and manipulation actions.

We also plan to study how to further extend input vocabularies by combining touch [10], [7], [56] and mid-air gestures [5] with physical objects [45], [65], [52] and classical input devices such as keyboards to enable users to input commands to the system or to involve other users in their workflow (request for help, delegation, communication of personal findings, etc.) [29], [51]. Gestures and objects encode a lot of information in their shape, dynamics and direction, that can be directly interpreted in relation with the user, independently from the display output. Physical objects can also greatly improve coordination among actors for, e.g., handling priorities or assigning specific roles.

OAK Project-Team

3. Research Program

3.1. Scalable and Expressive Techniques for the Semantic Web

The Semantic Web vision of a world-wide interconnected database of *facts*, describing *resources* by means of *semantics*, is coming within reach as the W3C's RDF (Resource Description Format) data model is gaining traction. The W3C Linking Open Data initiative has boosted the publication and interlinkage of a large number of datasets on the semantic web resulting to the Linked Open Data Cloud. These datasets of billions of RDF triples have been created and published online. Moreover, numerous datasets and vocabularies from different application domains are published nowadays as RDF graphs in order to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. RDF storage, querying, and reasoning is now supported by a host of tools whose scalability and expressive power vary widely. Unsurprisingly, some of the most scalable tools draw upon the existing models and architecture for managing structured data. However, such tools often ignore the semantic aspects that make RDF interesting. For what concerns the semantics, a delicate balance must be found between expressive power and the efficiency of the resulting data management algorithms.

- The team works on identifying tractable dialects of RDF, amenable to highly efficient query answering algorithms, taking into account both data and semantics.
- Another line of research investigates the usage of RDF data and semantics to help structure, organize, and enrich structured documents from social media. Based on such a rich model, we devised novel query answering algorithms which attempt to explore efficiently the rich social dataset in order to return the most pertinent answers to the users, from a social, structured and semantic perspective. This research is related to the DIGICOSME LabEx grant "Structured, Social and Semantic Search".
- To help users get acquainted with large and complex RDF graphs, we have started to work on an approach for RDF graph summarization: a graph summary is a smaller RDF graph, often by several orders of magnitude, which preserves the core structural information of the original graph and thus allows to reason about several important graph property on a much more manageable structure.

3.2. Massively Distributed Data Management Systems

Large and increasing data volumes have raised the need for distributed storage architectures. Among such architectures, computing in the cloud is an emerging paradigm massively adopted in many applications for the scalability, fault-tolerance and elasticity features it offers, which also allows for effortless deployment of distributed and parallel architectures. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment. For these reasons, cloud-based stores are an interesting avenue to explore for handling very large volumes of RDF data.

A recent development in this area is the start of our collaboration with social scientists from UNIV. PARIS-SUD, working on the management of innovation; we have started a collaborative research projects (ANR "Cloud-Based Organizational Design") where we perform an interdisciplinary analysis (both from a computing and from a business management perspective) on the adoption of cloud technologies within an enterprise.

3.3. Social Data Management and Crowdsourcing

The social Web blurs today the distinction between search, recommendation, and advertising (three paradigms for information access that have been so far considered mostly in separation). Our research in this area strives to find better adapted and scalable ways to answer information needs in the social Web, often by techniques at the intersection of databases, information retrieval, and data mining.

In particular, we study models and algorithms for personalized, or social-aware search in social applications. While progress has been made in this area, more remains to be done in order to address users' needs in practice, especially towards richer data models, and improving applicability and result relevance. For instance, when searching for tweets, their geographical location and recency may be as important for relevance as the textual and social aspects.

Furthermore, regarding quality of answers in response to searches, for various reasons (e.g., sparsity or tagging quality), meaningful results may often not be available. One response to this observation could be to turn to the crowd, the very users/publishers of the social media platform, and to turn this crowd into on-demand and query-driven sources of data. We study principled approaches for crowd selection (expert sourcing) and task assignment (data sourcing), in order to better answer ongoing social queries.

Beyond social links that represent just ties, a promising direction we also focus on in user-centric applications is to uncover implicit, potentially richer relationships from user interactions and to exploit them to improve core functionality such as search.

Moreover, we plan to investigate how crowdsourcing can be exploited to extract informations on user preferences, using techniques about noisy data management and provenance analysis.