Activity Report 2014

Section New Results

ATEAMS Project-Team

5. New Results

5.1. Highlights of the Year

- Davy Landman, Jurgen Vinju received a Best paper award nomination, for their paper "Empirical analysis of the relationship between CC and SLOC in a large corpus of Java methods"(ICSM'14).

5.2. Cyclomatic complexity ≠ Lines of Code

It has long been believed that cyclomatic complexity of source code correlates linearly with lines of code (SLOC). After extensive study of a large corpus of Java source code, Davy Landman and Jurgen Vinju refuted this belief [34]. This provides a new landmark in how to assess and measure the quality of software. In short: cyclomatic complexity measures something different than lines of code.

5.3. Language-Parametric, Capture-Avoiding Program Transformation

Hygienic transformations are well-studied in the area of programming languages that feature (syntax) macros. For instance, in Scheme, macro expansion is guaranteed to not involuntarily capture existing bindings, or allow new bindings to be captured. Together with Sebastian Erdweg and Yi Dai, Tijs van der Storm designed a technique, "name-fix", that can be used to ensure hygiene in arbitrary program transformations, even when source and target language are completely different [24].

5.4. Memory Efficient Hash Tries

The hash trie data structure is a common part in standard collection libraries of JVM programming languages such as Clojure and Scala. It enables fast immutable implementations of maps, sets, and vectors, but it requires considerably more memory than an equivalent array-based data structure. Michael Steindorfer designed a product family of hash tries to generate specialized Java source code [29]. A preliminary experiment on the implementation of sets and maps shows that this technique leads to a median decrease of 55% in memory footprint for maps and 78% for sets.

5.5. Reflection without Remorse

A series of list appends or monadic binds for many monads performs algorithmically worse when it is left-associated. Continuation-passing style (CPS) is well-known to cure this severe dependence of performance on the association pattern. The advantage of CPS dwindles or disappears if we have to examine or modify the intermediate result of a series of appends or binds, before continuing the series. Such examination is frequently needed, for example, to control search in non-determinism monads. Atze van der Ploeg (together with Oleg Kiselyov) developed an alternative approach that is just as general as CPS but more robust: it makes series of binds and other such operations efficient regardless of the association pattern [30]. This solution solves previously undocumented, severe performance problems in iteratees, LogicT transformers, free monads and extensible effects.

5.6. General Parser Combinators

Parser combinators are a well-known approach to parsing where grammars are represented using (higher-order) functions. Unfortunately, parser combinators are commonly implemented using recursive descent parsing as the underlying algorithm. As a result, most parser combinators frameworks do not support left-recursive rules, and may exhibit exponential runtime performance due to backtracking. Anastasia Izmaylova and Ali Afroozeh developed "general parser combinators" (GPC) which do not suffer from these problems: all context-free grammars are supported (even ambiguous ones) and performance is worst-case cubic. As result, GPC combines the expressiveness and performance guarantees of general parsing algorithms like GLL and GLR with the flexibility and extensibility of parser combinators.

BONSAI Project-Team

6. New Results

6.1. Highlights of the Year

- Amandine Perrin received the best paper award and the best oral presentation at the ISCB-LA 2014 international conference for the work on reconstruction of ancestral gene orders.

- Hélène Touzet was invited as a keynote speaker at the ALGO 2014 international conference. The topic of the talk was RNA bioinformatics.

BEST PAPERS AWARDS :

[7] **ISCB-Latin America**. A. PERRIN, J.-S. VARRÉ, S. BLANQUART, A. OUANGRAOUA.

6.2. High-throughtput sequence processing

- **Analysis of immunological rearrangements for leukemia diagnosis and monitoring.** High-throughput sequencing is spreading in the hospitals and many classical routines are now being transferred to this new technology. However in the specific case of lymphocyte monitoring, some complications arise. Classical bioinformatics software tools do not apply to the specificity of lymphocyte rearrangements. That is why we developed the software Vidjil (see 5.2) together with Lille hospital. This work has been published [5] and was also presented, as a poster, during the annual conference of the American Society of Hematology (ASH) [13]. We are now members of the EuroClonality-NGS work group which aims at providing a standardized way of monitoring leukemia using high-throughput sequencing at the European level.

- **New seeds for approximate pattern matching.** We addressed the problem of approximate pattern matching using the Levenshtein distance. Given a text T and a pattern P, find all locations in T that differ by at most k errors from P. For that purpose, we proposed a filtration algorithm that is based on a novel type of seeds, combining exact parts and parts with a fixed number of errors, that we called 01^*0 seed. Implementation has been performed on a Burrows-Wheeler transform. Experimental tests show that the method is specifically well-suited to search for short patterns (< 50 letters) on a small alphabet (*e.g.* DNA alphabet) with a medium to high error-rate (7 %–15 %). This work has been published in [9], and has a large number of applications in computational biology, such as finding microRNA targets, for example.

- **Spaced seeds and Transition seeds.** This year, two collaborative works have been published on the topic of spaced seeds and derivated models. The first work, resulting from a collaboration with Martin C. Frith from the *Computational Biology Research Center* (Tokyo), increases the sensivity of several search tools (among them, LAST, LASTZ, YASS,...) by computing specific seeds adapted to transition ratios observed during Eucaryotic comparisons. This work has been published in [3], together with the design of seeds obtained. The second work, issued from collaboration with Donald E.K. Martin from the *Department of Statistics* of the *North Carolina State University* (Raleigh), deals with the coverage of spaced seeds and shows how this criterion helps selecting good seeds for SVM string-kernels and alignment-free distances. This work has been published in [6].

6.3. RNA algorithms

- **A universal framework for RNA algorithms.** We have proposed a new generic specification framework, called *inverted coupled rewrite systems* that can deal with optimization problems on strings, trees, and arc-annotated sequences. It is specifically well-suited to handle RNA algorithms, such as alignment or folding algorithms. It is based on the following ideas. The solutions of combinatorial optimization problems are the inverse image of a term rewrite relation that reduces

problem solutions to problem inputs. A tree grammar is used to further refine the search space, and optimization objectives are specified as interpretations of these terms. All these constituents provide a mathematically precise and complete problem specification, leading to concise yet translucent specifications of dynamic programming algorithms. This work is a collaborative project with R. Giegerich from Universität Bielefeld, and has been published in [4].

- **RNA multistructures.** In many RNA families, the signature of the family cannot be characterized by a single consensus structure, and is mainly described by a set of alternate secondary structures. For example, certain classes of RNAs adopt at least two distinct stable folding states to carry out their function. This is the case of riboswitches, that undergo structural changes upon binding with other molecules, and recently some other RNA regulators were proven to show evolutionary evidence for alternative structure. The necessity to take into account multiple structures also arises when modeling an RNA family with some structural variation across species, or when it comes to work with a set of predicted suboptimal foldings. In this perspective, we have introduced the concept of RNA multistructures, that is a formal grammar based framework specifically designed to model a set of alternate RNA secondary structures. We provide several motivating examples and propose an efficient algorithm to search for RNA multistructures within a genomic sequence. This work was published in [8].

6.4. Ancestral gene order reconstruction

- In the field of **genomic rearrangement**, a topic of interest is to infer ancestral gene order from gene order known in extant species. The problem resumes to compute a set ancestral CARs (continuous ancestral regions) at a given node of a phylogeny. We designed a progressive homology-based method which iteratively detects and assembles ancestral adjacencies while allowing some micro-rearrangements of synteny blocks at the extremities of the progressively assembled CARs. Comparing with other methods we are able to produce more robust CARs with a very simple and efficient method. This work was published in [7].

6.5. Nonribosomal peptides

- **Monomeric structure.** The algorithm that identifies the monomeric structure of a polymer from its chemical structure has been finished and named s2m. It is based on a double index: A partial index constructed on the monomer database that uses a markovian model to speed up the search time ; and an index constructed on the fly on the studied polymer. This strategy was originally developed for nonribosomal peptides, but can be applied to any polymer.

- **Florine: Nonribosomal peptide synthetase annotations.** Florine [2] is a workflow dedicated to the discovery of new nonribosomal peptide synthetases. It describes sequential steps starting from DNA sequences leading to the design of candidate bioactive peptides. It is a useful tool for new drug discovery because it can be applied whatever the producing micro-organisms as it takes into account the enzymatic specificities related to each genus. This work was performed in collaboration with members of EPI Orpailleur (CRI Nancy Grand Est), Marie-Dominique Devignes and Malika Smaïl-Tabbone.

- **Activity prediction of small molecules.** Bayesian Belief Network was used for the first time to classify compounds according to their biological activity [1]. This method was applied on nonribosomal peptides and gave promising results on predicting their activity.

DOLPHIN Project-Team

6. New Results

6.1. Highlights of the Year

In [23], we have revisited the design and implementation of the Branch and Bound algorithm for solving on large scale distributed environments challenging permutation-based optimization problems such as Q3AP. The new approach includes original ways to efficiently deal with some crucial issues mainly, dynamic adaptive load balancing and fault tolerance. The approach allowed to solve to optimality for the first time a difficult Q3AP instance (Nug15) on the nation-wide Grid'5000 computational grid. The resolution was completed within less than 12 days using an average of 1,123 processing cores distributed over 6 Grid'5000 sites and peaked at 3,427.

6.2. Fitness Landscape Analysis for Multiobjective Optimization

Participant: Arnaud Liefooghe.

The properties of local optimal solutions in multi-objective combinatorial optimization problems are crucial for the effectiveness of local search algorithms, particularly when these algorithms are based on Pareto dominance. Such local search algorithms typically return a set of mutually non-dominated Pareto local optimal (PLO) solutions, that is, a PLO-set. In [34], we investigate two aspects of PLO-sets by means of experiments with Pareto local search (PLS). First, we examine the impact of several problem characteristics on the properties of PLO-sets for multi-objective NK-landscapes with correlated objectives. In particular, we report that either increasing the number of objectives or decreasing the correlation between objectives leads to an exponential increment on the size of PLO-sets, whereas the variable correlation has only a minor effect. Second, we study the running time and the quality reached when using bounding archiving methods to limit the size of the archive handled by PLS, and thus, the maximum size of the PLO-set found. We argue that there is a clear relationship between the running time of PLS and the difficulty of a problem instance.

Complementarily, in [25] we study the behavior of three elitist multi- and many-objective evolutionary algorithms in generating a high-resolution approximation of the Pareto set. Several search-assessment indicators are defined to trace the dynamics of survival selection and measure the ability to simultaneously keep optimal solutions and discover new ones under different population sizes, set as a fraction of the Pareto set size. Our study clarifies the ability and efficiency of the algorithms assuming scenarios where it is relatively easy to hit the Pareto set, showing the importance to properly assess algorithm's performance according to the task of the optimizer in many-objective optimization.

6.3. Combining dynamic programming and metaheuristics for the Unit Commitment Problem

Participants: Sophie Jacquin, Laetitia Jourdan, El-Ghazali Talbi.

DYNAMOP (DYNAmic programming using Metaheuristic for Optimization Problems) is a new dynamic programming based on genetic algorithm. It uses a representation based on a path in the graph of states of dynamic programming which is adapted to dynamic structure of the problem and facilitates the hybridization between evolutionary algorithms and dynamic programming. Experiments indicate that the proposed approach outperforms the best known in literature [44].

6.4. Multi-decoding strategy for Multi-objective Unit Commitment Problem

Participants: Sophie Jacquin, Lucien Mousin, Igor Machado, Laetitia Jourdan, El-Ghazali Talbi.

In the multiobjective version of the UCP taking the emission of gas into account, the dispatching problem remains easy to solve whereas considering it separatly remains interesting. A multi-objective GA handling binary vectors is applied. However for a binary representation there is a set of solutions of the dispatching problem that are pareto equivalent. In this approach a genotypic solution is associated with a set of phenotypic solutions. This set of solutions is from the optimal pareto front solution of the dispatching problem associated with the genotypic solution. As many phenotypic solutions are attached to a single genotypic solution, the fitness assignment and diversity assignment methods of NSGA-II have to be adapted. The multi decoding embedded approach has shown very good performances in comparison to two other less complex decoding systems.

6.5. Decomposition-Based Algorithms for Multiobjective Optimization

Participants: Dimo Brockhoff, Bilel Derbel, Arnaud Liefooghe, Gauvain Marquet, El-Ghazali Talbi.

Recently, there has been a renewed interest in decomposition-based approaches for evolutionary multiobjective optimization. Those algorithms decompose a multiobjective optimization problem into several single-objective optimization problems by using so-called scalarizing functions which are then simultaneously optimized by single-objective algorithms in a cooperative manner.

Our contributions to decomposition-based algorithms in 2014 has been three-fold. Firstly, we investigated in [28] the general impact of different scalarizing functions and their parameters on the search performance. We thereby abstracted from any specific algorithm and only considered the difficulty of the single scalarized problems in terms of the search ability of a (1+lambda)-EA on bi-objective NK-landscapes. Secondly, in [16], we proposed a new distributed heuristic for approximating the Pareto set of bi-objective optimization problems. Given a number of computing nodes, we self-coordinate them locally, in order to cooperatively search different regions of the Pareto front. As local information, every node uses only the positions of its neighbors in the objective space and evolves its local solution adaptively, based on what we term a 'localized fitness function'. We deployed our distributed algorithm using a computer cluster of hundreds of cores. At last, we enhanced the algorithm MOEA/D, a prominent example of a decomposition-based algorithm from the literature, by investigating the idea of evolving the whole population simultaneously at once. We thereby proposed new alternative selection and replacement strategies that can be combined in different ways within a generic and problem-independent framework [36].

6.6. Link-Heterogeneous work stealing for Branch-and-Bound Algorithms

Participants: T-T Vu, Bilel Derbel.

In this work [41], we push forward the design of parallel and distributed optimization algorithms running on link-heterogeneous systems where network latencies can deeply impact performance. We consider parallel Branch-and-Bound (B&B), viewed as a generic algorithm searching in a dynamic tree representing a set of candidate solutions built dynamically. A major challenge is then to deal with the irregularity of B&B computations and to distribute workload evenly at runtime. In this context, the random work-stealing paradigm has been proved to be extremely beneficial. However, it is known to perform loosely in non-homogeneous distributed systems where communications costs are a major obstacle for high performance. We there-by investigate the design of an effective work-stealing protocol dealing with the heterogeneity of network link latencies. We propose a generic distributed algorithm which can be easily implemented to fit different types of heterogeneity. The proposed algorithm extends on reference approaches, namely Probabilistic Work Stealing (PWS), and Adaptive Cluster-aware Random Stealing (ACRS); by introducing new adaptive control operations that are shown to be highly accurate in increasing work locality and decreasing steals cost. Through emulations on top of a real test-bed, we provide a comprehensive experimental analysis including: (i) a comparative study on a broad range of harsh network scenarios going from flat networks to more hierarchical grid-like networks, and (ii) an in-depth analysis of protocols' behavior at the aim of gaining new insights into dynamic load-balancing in heterogeneous distributed environments. Over all experimented configurations, our results show that although the proposed protocol is not tailored for a specific networked platform, it can save 30% execution time in average compared to its competitors, while demonstrating high quality self-adjusting capabilities.

6.7. New data structure for solving large permutation problems using multi-core B&B

Participants: Rudi Leroy, Nouredine Melab.

Solving large permutation problems using parallel B&B algorithms results in the generation of a very large pool of subproblems. Defining an efficient data structure is highly required to store and manage efficiently that pool. In [31], we have proposed a new dedicated data structure called *Integer-Vector-Matrix* (or *IVM* and redefined the operators of the B&B algorithm acting on it. We have also revisited the Work Stealing mechanism on multi-core processors. In the proposed approach, work units are coded in a coalesced way using factoradic-based intervals, and private IVMs are used to store and explore locally subsets of subproblems. The IVM-based approach has been experimented and compared to the approach based on concurrent linked-list, which is often used. The results show that our approach is more efficient in terms of memory usage and management time. In [31], we have investigated various work stealing strategies based on different victim selection and granularity policies. This later paper has been selected for a special issue in the CCPE international journal.

6.8. B&BGrid revisited for solving challenging Q3AP instances on large volatile computational environments

Participants: Nouredine Melab, El-Ghazali Talbi.

We have revisited the design and implementation of parallel B&B algorithms on multi-core (collaboration with UMONS, Belgium) and grid-wide environments (collaboration with University of Luxembourg and UMONS, Belgium)) for solving to otimality and efficiently large permutation problem instances. We have proposed a gridification approach of the B&B algorithm called B&B@Grid. This later includes a dynamic load balancing technique and a checkpointing mechanism for permutation problems. The approach has been validated through single-permutation Flowshop problem. In [23], we have extended the approach to deal with more than one permutation. To do that, we have revisited the design and implementation of the dynamic load balancing and checkpointing mechanisms for multiple permutation-problems. The new approach allowed the optimal resolution on a nation-wide grid (Grid'5000) of a difficult instance of the 3D quadratic assignment problem (Q3AP). To solve the instance, an average of 1,123 processing cores were used during less than 12 days with a peak of around 3,427 CPU cores.

DREAMPAL Team

5. New Results

5.1. Highlights of the Year

The papers [4] and [6] are published in journals (Software Testing Verification and Analisys, resp. Formal Aspects of Computing) that are among the best in their respective fields.

5.2. HoMade

HOMADE V5 is available from 03/2014. New features cover :

- new pipeline architecture with delayed conditional branch
- new unified FSM: Pipeline 2 stages
- renumbering of some IPs
- new activity management on the Slaves in 1D / 2D : by the master OnX , OnY, OnXY, and by the slaves the IPsleep removes the Slave from the next SPMDcall
- new bit per bit loading of program memories, for master and slaves
- new names for some components.
- new versions of a lot of IPs (inside)
- new communication network between Slaves: 2D torus ring with broadcast and communication on x or y axis
- new input binary file format (to respect !!)
- new test_bench for fast reading of instruction files
- new UART wrapper
- new assembler Hasm for those that do not speak binary
- nexys3 version for cheap platform experimentation (does not support more than 2x1 Slaves)
- V6 V7 xilinx supports up to 12 x 12 slaves
- Isim supports many more slaves !!!

More details can be found on www.lifl.fr/~dekeyser/Homade.

5.3. HiHope : A higher level language for the HoMade processor

HiHope is a programming language inspired by Forth used to program the HoMade processor. It includes language constructs for switching at runtime between hardware functions (implemented by IPs) and software functions in a transparent way. We also propose the notion of parallel function language construct. As a result, HiHope programs can use either hardware IPs or software functions, and can perform both sequential and parallel function calls, as well as sequential and parallel function redefinitions.

5.4. Integrating Profiling into MDE Compilers

This work [3] aims at improving performance by returning to the high-level models, specific execution data from a profiling tool enhanced by smart advices computed by an analysis engine. In order to keep the link between execution and model, the process is based on a traceability mechanism. Once the model is automatically annotated, it can be re-factored aiming better performances on the re-generated code. Hence, this work allows keeping coherence between model and code without forgetting to harness the power of parallel architectures. The example uses a transformation chain from UML-MARTE models to OpenCL code.

5.5. Language-Independent Symbolic Execution, Program Equivalence, and Program Verification

A significant part of our research project consists in applying formal techniques for symbolically executing and formally verifying HiHope programs, as well as for formally proving the equivalence of HiHope programs with the corresponding HoMade assembly and machine-code programs obtained by compilation of HiHope.

- Symbolic execution will detect bugs (e.g., stack underflow) in HiHope programs. Additionaly, symbolic execution is the natural execution manner of HiHope programs as soon as they contain (typically, underspecified) hardware IPs;

- program verification will guarantee the absence of bugs (with respect to specified properties, e.g., no stack underflow, no invocation of unavailable IPs, ...);

- program equivalence will guarantee that such above-mentioned bugs are also absent from the HoMade assembly and machine-code programs obtained by compilation of HiHope source code.

Since these languages are still evolving we decided to work (together with our colleagues from Univ. Iasi, Romania) on language-independent symbolic execution, program-equivalence, and program-verification techniques. In this way, when all the languages in our project become stable, we will be readily able to instantiate the above generic techniques on (the K formal definitions of) the languages in question. We note that all the techniques described below are also independent of K: they are applicable to other language-definition frameworks that use similar rewriting-based formal operational semantics.

5.5.1. Symbolic Execution

In [15] we propose a language-independent symbolic execution framework. The approach is parameterised by a language definition, which consists of a signature for the language's syntax and execution infrastructure, a model interpreting the signature, and rewrite rules for the language's operational semantics. Then, symbolic execution amounts to performing a so-called symbolic rewriting, which consists in changing both the model and the manner in which the operational semantics rules are applied. We prove that the symbolic execution thus defined has the properties naturally expected from it. A prototype implementation of our approach was developed in the K Framework. We demonstrate the genericity of our tool by instantiating it on several languages, and show how it can be used for the symbolic execution, bounded model checking, and deductive verification of several programs. With respect to earlier versions of this work, we have redefined symbolic execution in a more generic way and have included applications to model checking and deductive verification. The current version of the report [15] is submitted to a journal and is based on Andrai Arusoaie's PhD thesis [1], defended in September 2014 at Univ. Iasi (Romania). Andrei was co-supervised by Vlad Rusu and has since joined Dreampal as a postdoc.

5.5.2. Program Equivalences

In [6] we propose a logic and a deductive system for stating and automatically proving the equivalence of programs written in languages having a rewriting-based operational semantics. The chosen equivalence is parametric in a so-called observation relation, and it says that two programs satisfying the observation relation will inevitably be, in the future, in the observation relation again. This notion of equivalence generalises several well-known equivalences and is appropriate for deterministic (or, at least, for confluent) programs. The deductive system is circular in nature and is proved sound and weakly complete; together, these results say that, when it terminates, our system correctly solves the given program-equivalence problem. We show that our approach is suitable for proving equivalence for terminating and non-terminating programs as well as for concrete and symbolic programs. The latter are programs in which some statements or expressions are symbolic variables. By proving the equivalence between symbolic programs, one proves the equivalence of (infinitely) many concrete programs obtained by replacing the variables by concrete statements or expressions. The approach is illustrated by proving program equivalence in two languages from different programming paradigms. The examples in the paper, as well as other examples, can be checked using an online tool. This work was started in 2012. With respect to earlier versions, the new journal publication [6] includes a new and more general presentation of program equivalence as a temporal-logic formula, the generalisation of the

approach to nondeterministic-confluent language semantics, substantially more compact proofs, and a new application to corecursive programs.

In another work [10] we deal with a different kind of equivalence: *mutual equivalence*, which says that two programs are mutually equivalent if they both diverge or they end up in similar states. Mutual equivalence is an adequate notion of equivalence for programs written in deterministic languages. It is useful in many contexts, such as capturing the correctness of, program transformations within the same language, or capturing the correctness of compilers between two different languages. In the case of different languages one needs an operation called *language aggregation*, which we present in [11] in more detail, that combine two languages into a single one. We introduce a language-independent proof system for mutual equivalence, which is parametric in the operational semantics of two languages and in a state-similarity relation. The proof system is sound: if it terminates then it establishes the mutual equivalence of the programs given to it as input. We illustrate it on two programs in two different languages (an imperative one and a functional one), that both compute the Collatz sequence.

5.5.3. *Program Verification*

In [16] we present an automatic, language-independent program verification approach and prototype tool based on symbolic execution. The program-specification formalism we consider is Reachability Logic, a language-independent alternative to Hoare logics. Reachability Logic has a sound and relatively complete deduction system that offers a lot of freedom to the user regarding the manner and order of rule application, but it lacks a strategy for automatic proof construction. Hence, we propose a procedure for proof construction, in which symbolic execution plays a major role. We prove that, under reasonable conditions on its inputs (the operational semantics of a programming language, and a specification of a program, both given as sets of Reachability Logic formulas) our procedure is partially correct: if it terminates it correctly answers (positively or negatively) to the question of whether the given program specification holds when executing the program according to the given semantics. Termination, of course, cannot be guaranteed, since program-verification is an undecidable problem; but it does happen if the provided set of goals includes enough information in order to be circularly provable (using each other as hypotheses). We introduce a prototype program-verification tool implementing our procedure in the K language-definition framework, and illustrate it by verifying nontrivial programs written in languages defined in K. With respect to earlier versions of this work from 2013, program verification is now presented as a procedure (instead of a proof system), which leads to a direct implementation in the new version of our prototype tool. We also have a new theoretical result: *weak completeness*, which says that a negative answers returned by the verification procedure imply the fact that that the program does not meet its specification. Finally, since Andrei Arusoaie's arrival in the Dreampal team as a postdoc (Nov 2014) we have started working on certifying our verification procedure in the Coq proof assistant.

5.5.4. *Language Definitions as Rewrite Theories*

In [8] we study the relationships between language definition frameworks (e.g., the K framework) and rewrite theories (e.g., as those embodied in the Maude tool). K is a formal framework for defining the operational semantics of programming languages. It includes software tools for compiling K language definitions to Maude rewrite theories, for executing programs in the defined languages based on the Maude rewriting engine, and for analyzing programs by adapting various Maude analysis tools. A recent extension to the K tool suite is an automatic transformation of language definitions that enables the symbolic execution of programs, i.e., the execution of programs with symbolic inputs. In this paper we investigate more particularly the theoretical relationships between K language definitions and their translations to Maude, between symbolic extensions of K definitions and their Maude encodings, and how the relations between K definitions and their symbolic extensions are reflected on their respective representations in Maude. These results show, in particular, how analyses performed with Maude tools can be formally lifted up to the original language definitions. The results presented in this paper provide the theoretical underpinnings for the current version of the K-Maude tool.

5.6. Hardware chain for partial reconfiguration

The cost overhead due to the use of a softcore processor (MicroBlaze) to drive dynamic reconfiguration led us to explore alternative solutions. The one we have adopted is the use of a dedicated hardware IP (that can be invoked by HoMade) to control and manage dynamic and partial reconfiguration. This approach has led us to develop a complete hardware chain for partial bitstreams reads and writes. The proposed architecture is based on an external memory controller (DDR3) whose role is to manage bitstreams transfers from and to the DDR. Bitstreams loading are managed by a HoMade instruction implemented in a dedicated IP that drives the ICAP interface to transfer data into the reconfigurable area through the physical ICAP. One of the most important performance criteria of dynamic and partial reconfiguration is the reconfiguration time, that we always try to reduce while taking into account the compromise cost / area, speed and power consumption. Preliminary results give a transfer rate exceeding 500 MB/s. Such a result is clearly promising, especially since our hardware reconfiguration chain is constructed to be easily adaptable to SPMD (multi HoMade) needing parallel partial reconfiguration. This work has been the subject of a first communication in the GDR / SOCSIP conference in Paris: 11, 12, 13 June 2014.

5.7. Generic pixel distribution for parallel video processing application

In the frame of the PhD thesis of Karim Ali, we exploited this year the usage of parallel architectures for real-time image/video processing applications. Our main concern was the data distribution according to the parallelism level and respecting real-time processing constraint. As a first step, we proposed a generic pixel distribution model to be used with different image/video applications. Several parameters in the model can be configured according to the required size of the distributed macro-block with the possibility to control the sliding step in both horizontal and vertical directions. We have implemented our architecture on the Xilinx Zynq ZC706 FPGA evaluation board for two applications: the video downscaler (1:16) and the convolution filter. The experimental results showed the low hardware cost of the solution and how flexible is the model to be configured for different distribution scenarios. The architecture and experimental results were published in a paper entitled "A Generic Pixel Distribution Architecture for Parallel Video Processing" at Reconfigurable computing and FPGA international conference (ReConFig) in December 2014, Cancun, Mexico [7].

As a next step, we will reduce the operating clock frequency to decrease the power consumption while increasing the number of processing elements in the parallel architecture to maintain the same performance results. In this way, we will obtain a set of different design points differ in (area, power, other factors) and the system will have the ability to adapt its structure by moving between different design points according to the available resources to keep the same performance measurements. Furthermore, we will target intelligent transportation system, specially dynamic obstacle detection and tracking for autonomous vehicle navigation in collaboration with NAVYA (http://navya-technology.com).

5.8. Massively Parallel Dynamically Reconfigurable Multi-FPGA

In the frame of the PhD thesis of Venkatasubramanian Viswanathan, we conceived and validated a massively parallel and dynamically reconfigurable execution model for next generation high performance embedded systems. We have designed a multi-FPGA platform in order to conceive the massively parallel dynamically reconfigurable execution model. We have used several IP cores developed during the first two years of my PhD in order to test and validate the proposed model. We have proposed a new parallel dynamic reconfiguration mechanism for our architecture. We use our parallel reconfiguration model to reconfigure a subset or several IPs in parallel. We have proposed a partial reconfiguration model for next generation 3D FPGAs well-traced on the execution model (SPMD) in order to reconfigure in parallel a subset of the computing nodes. Finally, we have used the PicoComputing platform as an example to validate our proposed execution and reconfiguration models.

In order to demonstrate various features of such an architecture, we have implemented a scalable distributed secure H.264 encoding application with a FMC based high-speed sFPDP (serial Front Panel Data Port) data acquisition protocol to capture RAW video data. The system has been implemented on 3 different FPGAs, respecting the SPMD execution model managing several input video sources in parallel. We have measured various performance metrics of the proposed massively parallel dynamically reconfigurable system

and demonstrated several benefits. This work is going to be published in the FPGA 2015 conference as a poster titled "A Parallel And Scalable Multi-FPGA based Architecture for High Performance Applications" [13].

Later an ICAP controller was setup for dynamic partial reconfiguration in order to swap IPs during runtime on a single FPGA. We have used this IP along with the parallel communication feature of the multi-FPGA architecture, in order to broadcast a partial bitstream to all FPGAs at the same time and to do a parallel DPR in several FPGAs, thus emulating the reconfiguration model for next generation 3D FPGAs. These results represent a conceptual proof for a massively parallel dynamically reconfigurable next generation embedded computers that will use 3D PFGAs and reconfigure several logic layers in parallel.

5.9. HoMade-based MPSoC

The goal of this work is to build an MPSoC based on HoMade. The aimed system is a completely dynamically reconfigurable system. This mean that both the processing elements (HoMade) and the interconnection network are dynamically reconfigurable. The basic block in this system developped here is the interconnection network. It is a MIN (Multistage Interconnection Network) that would utilize oversizing techniques in order to reconfigure the network depending on the traffic.

5.10. Communication-Computation Overlap in Massively Parallel System-on-Chip

The Synchronous Communication Asynchronous Computation (SCAC) model is an execution model dedicated to the Massively Parallel System-on-Chip. This model proposes a novel processing paradigm, teh communication-computation overlap [17]. This concept does not only consider the programming level but also the implementation level. Using a decoupled control structure, the synchronous communication control is performed independently of the asynchronous computation control. Separating these two control phases allows the programmer to define programming strategies that overlap communication by computation to decrease the execution time.

To achieve this communication-computation overlap in SCAC architecture while avoiding the centralized control, in addition to the master controller, we define a second hierarchical control level, namely the slave controllers. The concept of this dual ontrol structure departs from the centralized configuration and instead of a uni-processor master controlling a set of parallel Processing Elements (PEs), the master cooperates with a grid of parallel slave controllers which supervises the activities of cluster of PEs. Based on this decoupled control structure, the programmer can manage the master-slave program to overlap communication by computation phase. Therefore, the basic idea to implement this paradigm is to divide the principal program into small blocks of parallel instructions, called Slave Program (SP), and send these blocks to the activated PEs of the system. Then, according to a predefined mask, the slave controllers send the begin execution orders. In parallel to computation, the slave controllers manage the synchronous inter-node communication. Distinguish communication from computation needs the separation of these two phases in different blocks. This repartition should be provided at programming level. Then, the overlapped execution of these blocks will be done in parallel according to the program description.

The aim of these last works is to define a new paradigm of a communication-computation overlap in massively parallel System-on-Chip. This paradigm allows to decrease the execution time of parallel programs using specific strategies in the programming level and a partially decoupled control system in the hardware level. The difficulty of implementing this paradigm lies in the coordination between the programming level and the architecture designing level in order to hide the communication cost.

<p style="text-align: center; color: red;">**FUN Project-Team**</p>

5. New Results

5.1. Highlights of the Year

- Opening of the 256 M3 sensor nodes of the Lille's FIT IoT Lab platform.
- We have designed a novel single-based localization method, UNS, for accurate localization of mobile devices that only needs a small aperture array unlike all previous works. UNS is currently under patenting.
- We have provided a set of recognized contributions in the area of Smart Cities, re-thinking their architecture and break vertical silos between every network and application.

5.2. Routing in FUN

Participants: Valeria Loscri, Nathalie Mitton, Riccardo Petrolo.

According to a wide range of studies, IT should become a key facilitator in establishing primary education, reducing mortality and supporting commercial initiatives in Least Developed Countries (LDCs). The main barrier to the development of IT services in these regions is not only the lack of communication facilities, but also the lack of consistent information systems, security procedures, economic and legal support, as well as political commitment. In [3], [10], we propose the vision of an infrastructure-less data platform well suited for the development of innovative IT services in LDCs. We propose a participatory approach, where each individual implements a small subset of a complete information system thanks to highly secure, portable and low-cost personal devices as well as opportunistic networking, without the need of any form of infrastructure. We review the technical challenges that are specific to this approach. Relying on such an infrastructure, wireless routing must be opportunistic and take advantages of the availability of every infrastructure point when in range. Two different approaches depending on the available devices are presented in [20] and [2]. When partial positions of nodes are available, the system can take advantage of such knowledge to enhance the routing performance. This is what has been investigated in[12] where coordinates are used in an opportunistic fashion when available.

5.3. Self-organization

Participants: Natale Guzzo, Valeria Loscri, Nathalie Mitton.

Self-organization encompasses several mechanisms. This year, the FUN research group has contributed to specific aspects; topology importance and clustering.

5.3.1. *Impact of the topology*

Wireless Sensor Networks (WSN) are composed of constrained devices and deployed in unattended and hostile environments. Most papers presenting solutions for WSN evaluate their work over random topologies to highlight some of their "good" performances. They rarely study these behaviors over more than one topology. Yet, the topology used can greatly impact the routing performances. [13] presents a study of the impact of the network topology on algorithm performance in WSNs and illustrate it with the geographic routing. Geographic routing relies on node coordinates to route data packets from source to destination. We measure the impact of different network topologies from realistic ones to regular and very popular ones through extensive simulation and experimentation campaigns. We show that different topologies can lead to a difference of up to 25% on delivery ratio and average route length and more than 100% on energy costs.

5.3.2. Clustering

Clustering in wireless sensor networks is an efficient way to structure and organize the network. It aims to identify a subset of nodes within the network and bind it a leader (i.e. cluster-head). This latter becomes in charge of specific additional tasks like gathering data from all nodes in its cluster and sending them by using a longer range communication to a sink or a Base Station (BS) which may be far away from the monitoring area. Many algorithms proposed in the literature compute the routing process by clustering the network and by designing new election mechanisms in which the cluster-heads are chosen taking account of the remaining energy, the communication cost and the density of nodes. However, they do not consider the connectivity to the BS, and assume that all the nodes or only few prefixed nodes are able to directly communicate with it. We believe that this assumption is not suitable for many applications of WSN and to tackle this problem we propose CESAR [14], a multi-hop and energy-efficient routing protocol for large-scale WSN which includes a new cluster-head selection mechanism aware of the battery level and the connectivity to the BS. Furthermore, our solution employs an innovative hybrid approach to combine both clustering and on-demand techniques in order to provide an adaptive behavior for different dynamic topologies. Simulation results show that our solution outperforms in terms of energy consumption and data delivery other known routing algorithms in the literature. Note that CESAR is currently the object of two pending patents.

5.4. Controlled mobility based services

Participants: Emilio Compagnone, Valeria Loscri, Karen Miranda, Nathalie Mitton, Tahiry Razafindralambo, Dimitrios Zormpas, Jean Razafimandimby Anjalalaina.

Sensors have more and more functionality in terms of capture techniques, communication capabilities, processing capabilities and energy harvesting. Another interesting feature available on sensors is mobility. The FUN research group tries to exploit the controlled mobility of sensors to solve some known issues in wireless sensors networks regarding deployment or routing but also raises some new challenges regarding coverage optimization and energy harvesting.

5.4.1. Coverage

Wireless sensors are used to gather information from a field of interest. In order to capture all the events in this field, the sensors must be properly placed. When the sensors have motion capabilities such as robots, the deployment can be optimized. The use of controlled mobility raises some new challenges and opportunities in the field of wireless sensor networks. Milan Erdelj and Karen Miranda in [33] presents the advances in context. They provide a detailed literature review regarding the techniques behind controlled mobility in order to deploy or redeploy sensors. When the wireless sensors are mobile, it is possible to optimize the capture of information regarding their time and space evolution. This allows the sensors to focus on different zones of interest depending on the evolution of the observed events. Valeria Loscri, Enrico Natalizio and Nathalie Mitton present a performance evaluation of different algorithms for zone of interest coverage in [18]. Their work particularly focuses on providing a set of distributed version of a combined particle swarm optimization and virtual forces algorithm. The proposed algorithms and their evaluation show an high reactivity to changing events and targets. Energy is an important constraint in wireless sensor networks and message exchange is a functionality that drains huge amount of energy. Dimitrios Zorbas and Christos Douligeris in [30] present a low-overhead localized algorithm for the target coverage problem in wireless sensor networks. To tackle this problem they propose two variations of a localized algorithm with low communication complexity in term of message exchange. The results show a great improvement in terms of communication cost while achieving an adequate network lifetime.

5.4.2. Connectivity and performance

Information gathered by sensors are to be processed in a remote location. The transportation from the point where the raw data is generated (the sensor) and the data processing unit (sink or other infrastructure) relies on routing techniques. Routing is a fundamental functionality of a wireless sensors network. Nicolas Gouvy, Nathalie Mitton and David Simplot-Ryl in their book chapter [34] provide a review of the routing techniques

described in the literature. They highlight the challenges, main issues and future work direction in this domain and provide some important assumption and characteristics that should be kept in mind when designing routing protocols for wireless sensor networks. When route between a source and the destination of data does not exist or cannot be established, using a mobile router is a possible solution. Christos Katsikiotis, Dimitrios Zorbas and Periklis Chatzimisios in [15] propose an algorithm that restores connectivity by the use of mobile wireless router after a routing failure. They provide a fast mechanism to heal the network and restore connectivity between the network partitions. In their solution, a mobile wireless router finds the end points that should be re-connect and place itself in the correct position to restore the connectivity. Their solution shows a fast restoration process based on the implementation done on a real robotic platform.

5.4.3. *Energy suppliance*

Energy is an important constraint in static wireless sensor networks and even more important when sensors are mobile. However, when sensors have motion capabilities, they can use this ability to move toward a recharging point in order to increase the network operation. Dimitrios Zorbas and Tahiry Razafindralambo in [31] use the motion capability of sensors to provide an algorithm that allow the sensor to go to a recharging point while minimize the impact of their movement on the network operation such as portioning or data gathering. They provide theoretical bounds on the realisation of such operation and evaluate the average behaviour of their algorithm based on extensive simulations. Both results show a big improvement in terms of network lifetime extension compared to the case where no replacement is performed and to the case where rerouting is considered.

5.4.4. *Video-based applications*

Video Surveillance and Target Detection represent key components for many organizations in terms of safety and security protocols. The value of Video Surveillance has become more sophisticated and very accurate, by leveraging specific sensors able to detect motion, heat, etc. In [17], Valeria Loscri, Michele Magno and Rosario Surace show how the nodes of a sensor network can learn which is their best position based on a certain number of WebCams that need to be "woken-up" when a suspicious event is detected. The main purpose is to reduce power consumption, especially in the case of Video Surveillance, when the most of the time the power is wasted by doing nothing. On the other hand, Target Detection, namely determining whether or not a target object exists in a video frame, has grown significantly with the recent advances in embedded computing and sensors which have opened the possibility to realize smaller and low-cost autonomous systems. In [16], Valeria Loscri, Nathalie Mitton and Emilio Compagnone show the feasibility of low-cost embedded system for detection of objects based either on the shape or on the color.

5.5. Security

Participants: Valeria Loscri, Nathalie Mitton.

Security has been always a critical issue both for the users and providers of wireless communication systems. The definition of novel paradigms and innovative communication systems, such as the Internet of Things (IoT) and the nanocommunication systems, exacerbated the criticality of security and privacy factors. These latter aspects are faced in [23] and [5]. In [23], Riahi et al. face with the security issues related to the IoT paradigm, by taking into consideration that this paradigm enable daily objects to become active participants of everyday activities. They envisage the main challenges and propose solutions to address them. In [5], Valeria Loscri et al. analyze the innovative aspects that characterize the molecular communication paradigm, by proposing innovative and revolutionary methods that take into consideration the very limited available resources (i.e. we work at molecular level and then we cannot leverage on high processing and computing capabilities) and the very high criticality of the potential applications of similar systems (e.g. in-vivo applications).

5.6. RFID

Participants: Ibrahim Amadou, Nathalie Mitton.

Due to the dedicated short range communication feature of passive radio frequency identification (RFID) and the closest proximity operation of both tags and readers in a large-scale dynamic RFID system, when nearby readers simultaneously try to communicate with tags located within their interrogation range, serious interference problems may occur. Such interferences may cause signal collisions that lead to the reading throughput barrier and degrade the system performance. Although many efforts have been done to maximize the throughput by proposing protocols such as NFRA or more recently GDRA, which is compliant with the EPCglobal and ETSI EN 302 208 standards. However, the above protocols are based on unrealistic assumptions or require additional components with more control packet and perform worse in terms of collisions and latency, etc. In [9], we explore the use of some well-known Carrier Sense Multiple Access (CSMA) backoff algorithms to improve the existing CSMA-based reader-to-reader anti-collision protocol in dense RFID networks. Moreover, the proposals are compliant with the existing standards. We conduct extensive simulations and compare their performance with the well-known state-of-the-art protocols to show their performance under various criteria. We find that the proposals improvement are highly suitable for maximizing the throughput, efficiency and for minimizing both the collisions and coverage latency in dense RFID Systems.

5.7. VANET

Participant: Nathalie Mitton.

VANET (Vehicular Networks) is an arising kind of network which features specific functionalities and requirements especially in terms of delay.

[26] analyzes the information delivery delay for roadside unit deployment in an intermittently connected vehicular network. A mathematical model is developed to describe the relationship between the average information delivery delay and the distance between two neighbor RSUs (Road Side Unit) deployed along a road. The derived model considers a straight highway scenario where two RSUs are deployed at a distance without any direct connection and vehicles are sparsely distributed on the road with road condition information randomly generated between the two neighbor RSUs. Moreover, the model takes into account the vehicle speed, the vehicle density, the likelihood of an incident, and the distance between two RSUs. The effectiveness of the derived mathematical model is verified through simulation results. Given the delay requirement of some time-critical applications, this model can be used to estimate the maximum distance allowed between two neighbor RSUs, which can provide a reference basis for the deployment of RSUs in such scenarios.

Abstract–Broadcasting is an effective routing paradigm for data dissemination in vehicular ad hoc networks (VANETs). One concern that arises with broadcasting is the broadcast storm problem, which would cause node contentions and data collisions, and thus degrade the transmission efficiency of a network. [27] proposes a Dynamic trAnsmission delaY based broadcast (DAYcast) protocol for a VANET. To alleviate the effect of the broadcast storm and improve the transmission efficiency of the network, DAYcast only allows the effective neighbors of a source vehicle to broadcast a received data packet and the selection of the effective neighbors are based on the position information on the one-hop neighbors of the source vehicle. Meanwhile, it allows each effective neighbor to wait a certain transmission delay before it broadcasts a received packet. The transmission delay of an effective neighbor depends on the distance between the neighbor and the source vehicle, and the number of effective neighbors of the source vehicle. Simulation results show that DAYcast can effectively improve the network performance in terms of network reachability and the successful delivery ratio as compared with existing weighted p-persistence broadcasting (WPB) and slotted 1-persistence broadcasting (SPB).

5.8. Smart cities architecture

Participants: Valeria Loscri, Nathalie Mitton, Riccardo Petrolo, Nicola Zema.

Smart City represents one of the most promising and prominent Internet of Things (IoT) applications. In the last few years, indeed, smart city concept has played an important role in academic and industry fields, with the development and deployment of various middleware platforms. However, this expansion has followed distinct approaches creating, therefore, a fragmented scenario, in which different IoT ecosystems are not able to communicate between them. To fill this gap, there is a need to re-visit the smart city IoT semantic and offer a global common approach. In order to allow cities to share data across systems and coordinate processes across domains, it is essential to break these silos. A way to achieve the purpose is sensor virtualization, discovery and data restitution. This last year, the FUN team has lead several investigations in this direction.

We have looked at the heterogeneity of devices and network technologies under a different perspective by not perceiving it as a limitation but as a potential to increase the connectivity in a smart city [1]. We propose a new generation of network nodes, called stem nodes, based on the innovative idea of 'stemness', which pushes forward the well-known self-configuration and self-management concepts towards the idea of node mutation and evolution. We also deployed prototypes that demonstrate the stem-node architecture and basic operations in different hardware platforms of common communication devices (an Alix-based router, a laptop and a smartphone)

In [7], we illustrate semantic interoperability solutions for IoT systems. Based on these solutions, we describe how the FP7 VITAL project aims to bridge numerous silo IoT deployments in smart cities through repurposing and reusing sensors and data streams across multiple applications without carelessly compromising citizens' security and privacy. This approach holds the promise of increasing the Return-On-Investment (ROI), which is associated with the usually costly smart city infrastructures, through expanding the number and scope of potential applications.

To this purpose, [21] browses the semantic annotation of the sensors in the cloud, and innovative services can be implemented and considered by bridging Clouds and Internet of Things. Things-like semantic will be considered to perform the aggregation of heterogeneous resources by defining the Clouds of Things paradigm. We survey the smart city vision, providing information on the main requirements and highlighting the benefits of integrating different IoT ecosystems within the cloud under this new CoT vision. This paper also discusses relevant challenges in this research area.

Going further, we also presented [22] a first implementation of this federation: a federation of FIT IoT-LAB within OpenIoT. OpenIoT is a middleware that enables the collection of data streams from multiple heterogeneous geographically dispersed data sources, as well as their semantic unification and streaming with a cloud infrastructure. Future Internet of Things IoT-LAB (FIT IoT-LAB) provides a very large scale infrastructure facility suitable for testing small wireless sensor devices and heterogeneous communicating objects. The integration proposed represents a way to reduce the gap existing in the IoT fragmentation, and, moreover, allows users to develop smart city applications by interacting directly with sensors at different layers. We illustrate it trough a basic temperature monitoring application to show its efficiency.

So, once all city network and infrastructure are set at the same level thanks to the above mentioned approaches, they can go further and offer additional services. An example of them is navigation[11] as also described in "Localization" section. Another example is to make use of the urban bikes [19]. Indeed, besides the growing enthusiast provoked by bicycles in smart and green cities and the benefit for health they bring, there still exists some reluctance in using bikes because of safety, road state, weather, etc. To counter-balance these feelings, there is a need to better understand bicycle users habits, path, road utilization rate in order to improve the bicycle path quality. In this perspective,in this paper, we propose to deploy a set of mobile sensors on bicycles to gather this different data and to exploit them to make the bike easier and make people want to ride bicycles more often. Such a network will also be useful for several entities like city authorities for road maintenance and deployment, doctors and environment authorities, etc. Based on such a framework,we propose a first basis model that helps to dimension the network infrastructure and the kind of data to be real time gathered from bikes. More specifically, we present a theoretical model that computes the quantity of data a bike will be able to send along a travel and the quantity of data a base station should be able to absorb.We have based our study on real data to provide first numerical results and be able to draw some preliminary conclusions and open new research directions.

5.9. Localization

Participants: Ibrahim Amadou, Roudy Dagher, Nathalie Mitton, Roberto Quilez, Nicola Zema.

Navigate in or based on a wireless sensor network present many advantages but it is still an open issue. We have focused on two particular cases in which navigation or WSN-based localization is needed [32]. The former aspect considers that sensors need to be visited on-demand by a mobile sink to offload data. This mobile sink thus needs to locate the data source. The second aspect feature a mobile entity that is needed to be localized.

In a event-based WSN, where is necessary a prompt response in terms of data processing and offloading, a set of mobile flying sinks could be a good option for the role of autonomous data collectors. For those reasons in [28], we propose a distributed algorithm to independently and autonomously drive a mobile sink through the nodes of a WSN and we show its preferability over more classical routing approaches especially in the presence of a localized generation of large amount of information. Our result shows that, in the case of fairly complete coverage of the area where the nodes lie, it is possible to promptly notify a mobile sink about the presence of data to offload, drive it to the interested area and achieve interesting performances. [29] enhanced the previous approach by relaxing some GPS-use assumptions. We show that, under fairly common circumstances, it is possible to set the trajectory of the mobile sink and fulfill the offloading requests without the needs of additional equipment installed on nodes. We show how our system is preferable over more classical routing solutions especially in the presence of localized generation of large amounts of information.

[11] proposes Ubiquitous Navigation System (UNS), a WSN-based navigation system, which takes benefit from a WSN mesh deployment to provide a local navigation service. The positioning part of the system uses Angle of Arrival (AoA) measurements to estimate the vehicle position on the map. Based on a realistic network scenario, extracted from a city map using Google Maps, we study the performance of Triangulation using AoA in a smart urban environment that exhibits topology related constraints. Simulations results show that such constraints lead to particular spatial distribution of the anchor nodes that affects both positioning accuracy and beacon packets reception rate. We also propose and evaluate the use of the network communication range as a technique to mitigate the effect of geometric dilution of precision (GDOP). The simulation results show that this technique successfully detected GDOP-affected positions and thus significantly enhanced the positioning accuracy. One of the biggest strengths of UNS is that it relies on a single anchor unlike literature approaches. The different underlying studies are detailed in [38] in which we study the ambiguity of source localization using signal processing of large aperture antenna arrays under spherical wave propagation. This novel localization approach has been recently proposed, providing an estimate of the source position by means of two methods: geometrical and analytical. The former finds the source position as the estimate of circular loci, the latter as a solution of a linear system of equations. Although this method is proved to work for a general array geometry, we show that it suffers from ambiguities for a particular class of array geometries. Namely, in 2D, we prove that when the array geometry is linear or circular, there exist two possible solutions where only one corresponds to the actual position of the source. We also prove a relation of symmetry between the solutions with respect to the array geometry. This relation is very useful to assist the disambiguation process for discounting one of the estimates. By extension to 3D, planar (resp. spherical) arrays exhibit the same behavior i.e they provide two symmetrical estimates of the source position when the latter is not on the array plane (resp. sphere).
Note that UNS is currently a pending patent.

<p style="text-align:center">**LINKS Team**</p>

5. New Results

5.1. Highlights of the Year

In the objective Querying Heterogeneous Linked Data, Slawomir Staworko and Iovka Boneva have developed new ways to define schema for Graph Database and RDF [19]. This work has been influencing a group work of W3C on defining a schema for the DF format. This work is a continuation of [3] (by Iovka Boneva, Radu Ciucanu and Slawomir Staworko) developing a new schema for unordered trees over XML. Due to these works, Boneva is now a member of the Data Shapes Working Group which mission is to produce a language for defining structural constraints on RDF graphs. http://www.w3.org/2014/data-shapes/charter

In the objective Managing Dynamic Linked Data, the main break through is the development of QuixPath that now covers 100 per cent of the XPathMark, a W3C benchmark for the language Xpath (querying XML trees). In particular, it includes aggregation operators, joins and arithmetics operations. The core of QuixPath is based on techniques presented in [6] (by Tom Sebastian, Denis Debardieux and Joachim Niehren).

In the objective Linking Data Graphs, different methods have been developed to learn queries over graph. More precisely, the queries learned are conjunctive queries with joins. These techniques have been presented in [13] and demonstrated in [4] at the conference VLDB.

5.2. Querying Heterogeneous Linked Data

Angela Bonifati, Gianvito Summa, Esther Pacitt (U Montpellier 2)i and Fady Draidi (U Montpellier 2) [5] consider peer-to-peer data management systems (PDMS), where each peer maintains mappings between its schema and some acquaintances, along with social links with peer friends. In this context, the goal is reformulating conjunctive queries from a peer's schema into other peer's schemas. Precisely, queries against a peer node are rewritten into queries against other nodes using schema mappings thus obtaining query rewritings. They propose a new notion of 'relevance' of a query with respect to a mapping that encompasses both a local relevance (the relevance of the query w.r.t. the mapping) and a global relevance (the relevance of the query w.r.t. the entire network). Based on this notion, they conceived a new query reformulation approach for social PDMS which achieves great accuracy and flexibility. This has been implemented and experimented in a prototype.

Pierre Bourhis, Andreas Morak and Andreas Pieris [14] investigated classes of queries for which the problem of open query answering of disjunctive guarded TGDs a decent complexity (e, g exptime). The complete picture of the complexity of answering (unions of) conjunctive queries under the main guarded-based classes of disjunctive existential rules has been recently settled. It has been shown that the problem is very hard, namely 2ExpTime-complete, even for fixed sets of rules expressed in lightweight formalisms.The central objective of the present paper is to understand whether simpler query languages (bounded tree width and acyclic queries) have a positive impact on the complexity of query answering under the main guarded-based classes of disjunctive existential rules.

In [3], a new formalism for schema for unordered trees have been developed. It is based on a notion of regular expressions of multisets of labels. Different problems of static analysis like emptiness and containment are studied and their complexity. Different simpler schema are studied leading to interesting complexity for the different studied problems. Finally, they study the expressive power of the proposed schema languages and compare them with yardstick languages of unordered trees (FO, MSO, and Presburger constraints) and DTDs under commutative closure. The results show that the proposed schema languages are capable of expressing many practical languages of unordered trees and enjoy desirable computational properties.

In [7], Adrian Boiret, Vincent Hugot and Joachim Niehren and Ralf Treinen (University Paris 7) proposes a notion deterministic tree automata for unordered trees. While the existing notions are well-investigated concerning expressiveness, they all lack a proper notion of determinism, which makes it difficult to distinguish subclasses of automata for which problems such as inclusion, equivalence, and minimization can be solved efficiently. In this paper, the authors propose and investigate different notions of "horizontal determinism", starting from automata for unranked trees in which the horizontal evaluation is performed by finite state automata.

5.3. Managing Dynamic Linked Data

Tom Sebastian, Denis Debarbieux, Olivier Gauwin (U Bordeaux), Joachim Niehren, Mohamed Zergaoui (Innovimax) [6] present new techniques to evaluate XPath queries on trees received in a streaming way. It introduce early nested word automata in order to approximate earliest query answering algorithms for nested word automata. The notion early query answering algorithm is based on stack-and-state sharing for running early nested word automata on all answer candidates with on-the-fly determinization. These techniques allow to implement a more important part of Xpath and outcome all the previous tools in coverage of XpathMark benchmark.

5.4. Linking Data Graphs

Angela Bonifati, Radu Ciucanu, Slawomir Staworko developed techniques to learn conjunctive queries from example given by a user. The main part is to infer joins between relations from the positive and negative tuples. Different techniques to deduce informative examples are presented and interestingly they can be done in polynomial time. The techniques are published in [13] and demonstrated in [4].

Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Slawek Staworko, Marc Tommasi [16] explain how to learn sequential top-down tree-to- word transducers (STWs). First, they present a Myhill-Nerode characterization of the corresponding class of sequential tree-to-word transformations (STW). Next, they investigate what learning of stws means, identify fundamental obstacles, and propose a learning model with abstain. Finally, they present a polynomial learning algorithm

<div align="center" style="color:red">MAGNET Team</div>

6. New Results

6.1. Highlights of the Year

We developed a new framework for high order learning [4].

We have illustrated the usefulness of automatically annotated examples in complex learning supervised by few training examples [2], [1].

We propose a new algorithm for semi-supervised spectral clustering and apply it to the NLP task of noun phrase coreference resolution [6].

6.2. Higher-order Learning with Graphs

Along the thesis of THOMAS RICATTE, in [4] and [8], we propose methods for learning from interactions between groups in networks. We propose a proper extension of graphs, called hypernode graphs as a formal tool able to model group interactions. A hypernode graph is a collection of weighted relations between two groups of distinct nodes. Weights quantify the individual participation of nodes to a given relation. We define Laplacians and kernels for hypernode graphs and and prove that they strictly generalize over graph kernels and hypergraph kernels. We prove that hypernode graphs correspond to signed graphs such that the matrix $D - W$ is positive semidefinite. As a consequence, homophilic relations between groups may lead to non homophilic relations between individuals. We define the notion of connected hypernode graphs and a resistance distance for connected hypernode graphs. We propose spectral learning algorithms on hypernode graphs allowing to infer node ratings or node labelings. As a proof of concept, we model multiple players games with hypernode graphs and we define skill rating algorithms competitive with specialized algorithms.

6.3. Natural Language Processing

In [6] (presented by DAVID CHATEL at the ECML–PKDD and CAp'2014 conferences) we propose a new algorithm for semi-supervised spectral clustering and apply it to the task of noun phrase coreference resolution. The main insight is in the inclusion of pairwise constraints into spectral clustering: our algorithm learns a new representation space for the data together with a distance in this new space. The representation space is obtained through a constraint-driven linear transformation of a spectral embedding of the data, and constraints are expressed with a Gaussian function that locally reweights the similarities in the projected space. A global, non-convex optimization objective is then derived and the model is learned via gradient descent techniques. Our algorithm is evaluated on the CoNLL-2012 coreference resolution shared task dataset, and shows some encouraging results.

In [2] and [1], we develop a new approach for the automatic identification of so-called implicit discourse relations. Specifically, our system combines hand labeled examples and automatically annotated examples based on explicit relations using several simple methods inspired by work in domain adaptation. Our system is evaluated empirically on the Annodis corpus, a French corpus annotated with discourse structures. Our system yields significant performance gains compared to only using hand-labeled data or using only automatically annotated data.

6.4. Ongoing work

6.4.1. Adaptive Graph Construction

We worked on developing a new algorithm in order to construct a graph in a adaptive way for a specific task. More precisely, we looked for a metric learning algorithm that could depend on the target task. Previous works on metric learning ([12]) aim at learning a relevant metric using a linear approach, which cannot capture the non-linearity of the data. Our approach, instead, aims at learning the most appropriate non-linear data projection for the target task. For this purpose, we train a neural network with relative constraints depending on the target task and a target classic metric (e.g. euclidean distance, cosine similarity, ...), in order to make the metric meaningful for the new data representation and our target task.

6.4.2. Correlation Clustering and Similarity/Dissimilarity Links

From a mathematical point of view, signed networks are graphs whose edges carry a sign representing the positive or negative nature of the relationship between the incident nodes. These structures are extremely useful for modeling, at the same time, similarity and dissimilarity object relationships. Given an undirected signed graphs, in the Correlation Clustering problem the goal is to find a node partition into clusters minimizing the number of negative (dissimilarity) edges linking two nodes within the same cluster and the number of positive (similarity) edges between different clusters.

We focused on devising an algorithm able to solve the Correlation Clustering problem for general input signed graphs (if the input is a complete signed graph the problem is proven to be much easier). One of the main objective of this work is the use of the proposed algorithm for creating a learner able to predict the unknown edges signs of a given signed graph. This prediction task is known as Link Classification in signed graphs. In fact, given an undirected signed graph whose edge set is split into training and test set, we could use the Correlation Clustering solution working for general input graphs for partitioning the training set and using the node partition generated for predicting the test edge signs. Moreover, one could exploit such an algorithm for developing new strategies for the Link Classification problem operating within the online and active Machine Learning setting.

Since the node set partitioning turns out to be strictly related to the Link Classification problem, we also focused on the very challenging goal of obtaining a deep understanding of the complex interplay between Link and Node Classification. More precisely, we investigated the relationships between the Vapnik Chervonenkis dimension of any given set of hypothesis space of node and edge similarity functions operating within this framework.

6.4.3. Ranking from Pairwise Sets of User Preferences

Given a set of objects (vertices of a graph) and a set of pairwise preference labels between objects (directed edges connecting vertices) which may be non-transitive due to irrationality or arbitrary noise, what is a correct way to sample preference labels for ordering the set of objects? This long standing open problem is, as far as we know, unsolved when each pairwise preference labels refers to two (disjoint) sets of objects (vertices). This framework can be easily motivated considering that quite often, in many real world contexts, users express their preferences between sets of items rather than single items, and turns out to be strictly connected with our recent model of hypergraphs with bipartite hyperedges [4]. We are working on devising a new algorithm able to rank a given set of items (graph node set) when only comparisons between sets containing at least 2 items are allowed. This challenging and interested problem is, as far as we know, quite novel and can be studied within different Machine Learning setting (online, batch, active, ...). The preliminaries results we are obtaining, when setting the cardinality of the item sets equal to 2, are encouraging and indicate that it could be possible to extend our strategies in order to deal with larger item sets.

6.5. Other results

In this section we provide the results we obtained that are not related with our main research directions.

In [3] we study the problem of learning sequential top-down tree-to-word transducers (STWs). First, we present a Myhill-Nerode characterization of the corresponding class of sequential tree-to-word transformations (STW). Next, we investigate what learning of stws means, identify fundamental obstacles, and propose a learning model with abstain. Finally, we present a polynomial learning algorithm.

6. New Results

6.1. Highlights of the Year

A. Gloria, S. Neukamm, and F. Otto published their recent contribution [17] on quantitative homogenization in Inventiones Mathematicae.

As a plenary speaker of the World Congress of Computational Mechanics in Barcelone in July 2014, P. Le Tallec (Ecole polytechnique) presented our joint results [15], [25].

6.2. Quantitative stochastic homogenization

A. Gloria, S. Neukamm (Univ. Dresden), and F. Otto (MPI for mathematics in the sciences, Leipzig) developed in [17] a general approach to quantify ergodicity in stochastic homogenization of discrete elliptic equations. Using a parabolic approach, they obtained optimal estimates on the time-decay of the so-called environment seen from the particle. This allowed them to prove optimal bounds on the corrector gradient and the corrector itself in any dimension (thus improving on [4]). They also obtained the first error analysis of the popular periodization method to approximate the homogenized coefficients.

In [32], A. Gloria and F. Otto extended their results [4], [5] on discrete elliptic equations to the continuum setting. They treated in addition the case of non-symmetric coefficients, and obtained optimal estimates in all dimensions by the elliptic approach (whereas [4], [5] were suboptimal for $d = 2$).

In [28], A. Gloria and D. Marahrens (MPI for mathematics in the sciences, Leipzig) extended the annealed results [51] on the discrete Green function by D. Marahrens and F. Otto to the continuum setting. As a by-product of their result, they obtained new results in uncertainty quantification by estimating optimally the variance of the solution of an elliptic PDE whose coefficients are perturbed by some noise with short range of dependence.

In their recent work [29], A. Gloria, S. Neukamm, and F. Otto developed a regularity theory for random elliptic operators inspired by the contributions of Avellaneda and Lin [39] in the periodic setting and of our visitor S. Armstrong with C. Smart [38]. This allowed them to consider coefficients with arbitrarily slow decaying correlations in the form of a family of correlated Gaussian fields.

In [30], A. Gloria and J. Nolen (Duke Univ.) proved a quantitative central limit theorem for the effective conductance on the discrete torus. In particular, they quantified the Wasserstein distance between a normal random variable and the CLT-like rescaling of the difference between the approximation of the effective conductance by periodization and the effective conductance. Their estimate is sharp and shows that the Wasserstein distance goes to zero (up to logarithmic factors) as if the energy density of the corrector was iid (which it is not). This completes and settles the analysis started in [17] on the approximation of homogenized coefficients by periodization by characterizing the limiting law in addition to the scaling.

6.3. Derivation of nonlinear elasticity from polymer-physics

In [15], A. Gloria, P. Le Tallec (Mechanics department, Ecole polytechnique), and M. Vidrascu (Project-team REO, Inria) numerically investigated the nonlinear elasticity model obtained in [1] by discrete stochastic homogenization, and compared it to the standard measurements by Treloar on natural rubber. The implementation was realized in the Modulef software. These results are in rather good agreement, which shows that the approach seems to be promising.

In [25], M. de Buhan (CNRS, Univ. Paris Descartes), A. Gloria, P. Le Tallec and M. Vidrascu proposed a numerical method to produce analytical approximations (that can be used in practical nonlinear elasticity softwares) of the numerical approximations obtained in [15] of the discrete-to-continuum energy density derived theoretically in [1]. This numerical method is based on the parametrization of the set of polyconvex Ogden laws and on the combination of a least square method and a genetic algorithm (cf. CMA-ES).

6.4. Numerical homogenization

Inspired by the quantitative analysis of [17] and [48], Z. Habibi (former SIMPAF post-doctoral fellow) and A. Gloria introduced in [14] a general method to reduce the so-called resonance error in numerical homogenization, both at the levels of the approximation of the homogenized coefficients and of the correctors. This method significantly extends [2]. The method relies on the introduction of a massive term in the corrector equation and of a systematic use of Richardson extrapolation. In the three academic examples of heterogeneous coefficients (periodic, quasiperiodic, and Poisson random inclusions), the method yields optimal theoretical and empirical convergence rates, and outperforms most of the other existing methods.

6.5. Nonlinear Schrödinger equation

S. De Bièvre, S. Rota Nodari (CEMPI postdoc 2013-2015) and F. Genoud (CEMPI visitor, September 2013) have explained the geometry underlying the so-called energy-momentum method for proving orbital stability in infinite dimensional Hamiltonian systems. Applications include the orbital stability of solitons of the NLS and Manakov equations. This work is to appear as a chapter (120p) in the first volume of the CEMPI Lecture Notes in Mathematics, cf. [24].

6.6. Kicked rotors

S. De Bièvre and his PhD student E. Soret rigorously proved the growth rate of the energy in a Markovian model for stochastic acceleration of a particle in a random medium, cf. [34].

6.7. Time integration of Hamiltonian system with noise

G. Dujardin introduced an energy preserving method for Hamiltonian dynamics perturbed by a multiplicative noise, cf. [11].

6.8. Miscellaneous results

The MEPHYSTO team is currently hosting former members of the project-team SIMPAF who focus on numerical methods for dissipative systems:

- corrosion models [19], [23],
- fluid mechanics [9], [21], [27], [10],
- numerical analysis for asymptotic preserving properties [8], [7],
- a posteriori estimates [20].

T. Gallouët also made contributions in optimal transport [22], [26].

MINT Project-Team

6. New Results

6.1. Highlights of the Year

- "Adoiraccourcix : sélection de commandes sur écrans tactiles multi-points par identification des doigts" [31] received the *best paper award* from the IHM 2014 conference;
- "L'ordinateur portable comme instrument de musique" [41] received the *best demo award* from the IHM 2014 conference.

6.2. Impact of form factors and input conditions on absolute indirect-touch pointing tasks

Absolute indirect interaction maps the absolute position of a device's end-effector to the absolute position of a remote on screen object. Despite its long-time use with graphics tablets and growing use in research prototypes, little is known on the influence of form factors and input conditions on pointing performance with such a mapping. The input and display can have different sizes and aspect ratios, for example. The on-screen targets can vary in size. Users can look solely at the display or at the input device as well. They can also hold the input device in certain cases, or let it rest on a table. We ran two experiments designed to investigate the influence of all these factors on absolute indirect-touch pointing performance [20], [11].

The first experiment focused on input device size and input conditions and revealed that users get higher performance when they can look at the input surface (even if nothing is displayed on it). In addition we found that the smallest target size users can acquire in motor space is not constant across different input dimensions but degrades as the input size increases. The second experiment focused on scale effects and aspect ratio and revealed users' performance is not affected by scale but that aspect ratio matters: similar input and output aspect ratios lead to better performance. This findings led us to list four main recommendations for the design of touch input surfaces with applications supporting absolute direct interaction.

6.3. Direct and indirect multi-touch interaction on a wall display

Multi-touch wall displays allow to take advantage of co-located interaction (direct interaction) on very large surfaces. However interacting with content beyond arms' reach requires body movements, introducing fatigue and impacting performance. Interacting with distant content using a pointer can alleviate these problems but introduces legibility issues and loses the benefits of multi-touch interaction. We introduced WallPad [30], [11], a widget designed to quickly access remote content on wall displays while addressing legibility issues and supporting direct multi-touch interaction (Figure 1). To support multi-touch on such a wall display, we developed a custom system using front diffuse illumination and 4 cameras. Our system can detect 50+ simultaneous contacts with a precision between 3 and 5 mm.

6.4. Sketching dynamic and interactive illustrations

We collaborated with Autodesk Research in Toronto (as a scientific consultant) on a project whose focus was to design and develop tools that enable artists to bring life to illustrations with subtle, continuous animation effects and infusing interactive behavior to the drawings. We believe designers, artists and creators should be able to communicate with computers the way they think about art and animation. This motivated Autodesk to develop interfaces that facilitate powerful ways of thinking and content creation with freeform sketching and direct manipulation, thus offering an alternative to complex professional animation tools. Our design combines the complementary affordances of humans and computers by utilizing by-example phenomena, thus preserving expressiveness and personal style, yet reducing tedium.
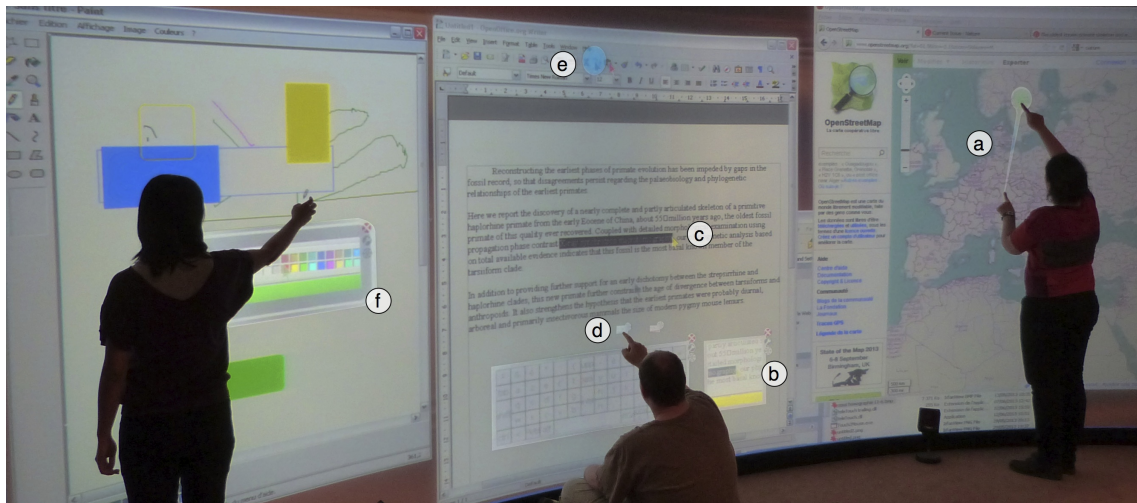
Figure 1. A user creating a WallPad (a). Another using two WallPads to select (b) some text (c) and click (d) on Copy/Paste buttons (e) beyond reach. A third user accessing a color palette located at the bottom of the screen through a WallPad (f).

The outcome of the collaboration is Kitty [23], a sketch-based tool for authoring dynamic and interactive illustrations (Figure 2). Artists can sketch animated drawings and textures to convey the living phenomena, and specify the functional relationship between its entities to characterize the dynamic behavior of systems and environments. An underlying graph model, customizable through sketching, captures the functional relationships between the visual, spatial, temporal or quantitative parameters of its entities. As the viewer interacts with the resulting dynamic interactive illustration, the parameters of the drawing change accordingly, depicting the dynamics and chain of causal effects within a scene. The generality of this framework makes our tool applicable for a variety of purposes, including technical illustrations, scientific explanation, infographics, medical illustrations, children's e-books, cartoon strips and beyond. A user study demonstrates the ease of usage, variety of applications, artistic expressiveness and creative possibilities of our tool.

Kitty is a follow up of a previous project, Draco   [50], a prototype sketch-based interface that allows artists and casual users alike to add a rich set of animation effects to their drawing, seemingly bringing illustrations to life such as a school of fish swimming, tree leaves blowing in the wind, or water rippling in a pond. Draco was realized before Fanny Chevalier joined Inria. Kitty is the result of a collaboration between Autodesk Research (inventor) and Inria (scientific consultant). A patent has been filed by Autodesk Research for Kitty, and the company is currently developping a commercial application based on the research protoype.

6.5. The not-so-staggering effect of staggered animations

Interactive visual applications often rely on animation to transition from one display state to another. There are multiple animation techniques to choose from, and it is not always clear which should produce the best visual correspondences between display elements. One major factor is whether the animation relies on staggering—an incremental delay in start times across the moving elements. It has been suggested that staggering may reduce occlusion, while also reducing display complexity and producing less overwhelming animations, though no empirical evidence has demonstrated these advantages. We empirically evaluated the effect of two staggering techniques on tracking tasks, focusing on cases that should most favour staggering [14]. We found that introducing staggering has a negligible, or even negative, impact on multiple object tracking performance. The potential benefits of staggering may be outweighed by strong costs: a loss of common-motion grouping
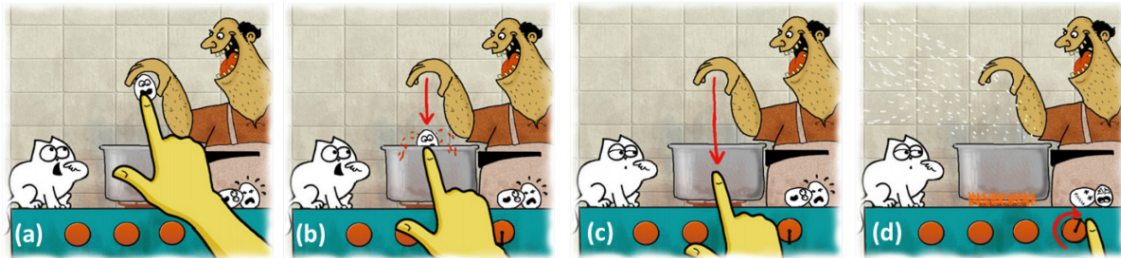
Figure 2. Example of a dynamic interactive illustration authored with Kitty. (a) Objects in the scene are interactive: the egg held by the cook can be dragged down, as if falling into the pot, triggering subsequent animations, such as soup splashes (b) and closing of the cat's eyelids (c). Turning the knob increases the fire and steam (d). The resulting dynamic illustration captures the living nature of the scene, where the gas stove flames burn and steam emits from the pot.

information about which objects travel in similar paths, and less predictability about when any specific object would begin to move.

6.6. Flexible contextual retrieval of chosen documents and windows

Users of Personal Computers interact with a large number of resources to do their work. To handle their different tasks, they need their documents to be readily available, and as the number of activities and documents increase, systems must offer adequate support for quick retrieval of these resources. The Hotkey Palette [29] is a quick retrieval facility that we designed that uses hotkeys and makes them visible and configurable through a quasi-modal always-available on-screen keyboard. This facility contributes to the state of the art in three ways. It extends on-screen keyboard interaction by providing feedback on the state of the linked resources, it provides persistent and integrated access to local windows and files and other online resources, and it provides flexible control over contextualization by leveraging existing resource hierarchies.

6.7. Multi-touch command selection using finger identification

Hotkeys are a critical factor of performance for expert users in WIMP interfaces. Multi-touch interfaces, by contrast, do not provide such efficient command shortcuts. Adoiraccourcix leverages finger identification to introduce quick command invocation integrated with direct manipulation in this context (Figure 3). We illustrated its use in a vectorial drawing application and ran preliminary user studies comparing it to classical user interfaces. Results suggest that once mastered, it provides very powerful means of interaction [31], [44], [43].

Figure 3. Partial illustration of the Adoiraccourcix' logics.

6.8. Impact of the localization and activation of mode switchers

Input devices have a limited number of buttons and degrees of freedom, but they are used to control many functionnalities. Modes and quasi-modes makes it possible to map several actions to the same input. For example keys of a keyboard either input a letter or trigger a command. Delimiters allow users to switch between the modes. On the keyboard, the default mode is often text entry and pressing the Ctrl key switches the mode to command mode. This choice was made at a time when the mouse was not widespread. In [33] , we explored the possibility to place mode switchers on the mouse and experimented the benefits. We showed that there is a performance benefit if the current tasks are essentially mouse-based. In particular we showed that using mode switchers on the mouse reduces homing the dominant hand between the mouse and the keyboard.

6.9. A serial Architecture for a collaborative robot

The haptic magnifier consists in using a serial architecture, where a motor is inserted between a tool and a user's hand (figure 4). By this way, the tool's speed v_o can be changed relatively to user's speed v_i, by controlling motor's speed. The haptic rendering of a load can then be changed, and fine details can be more easily detected.
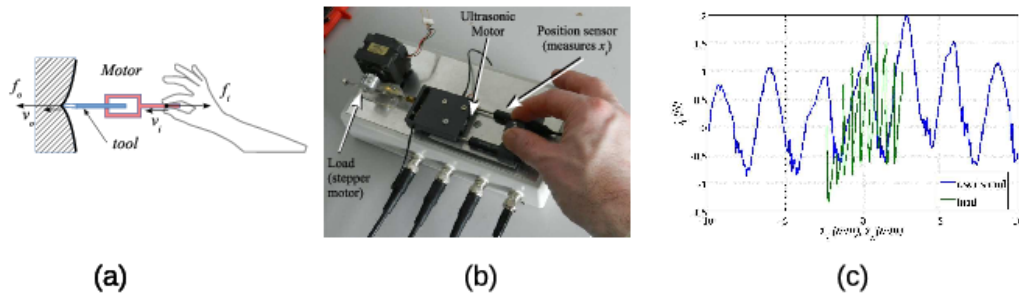

Figure 4. The Haptic Magnifier; (a) the serial architecture with a motor inserted to achieve a haptic magnifier, (b) the implementation with an ultrasonic Motor, and (c) the resulting rendering at load's end and user's end.

The haptic magnifier is built up with an ultrasonic motor, whose characteristic is low speed - high torque. So the tool and the end-effector can be directly connected to the motor, leading to a lightweight architecture. The user's study presented in [21] have shown that the precision in using the tool could be improved during a freehand manipulation.

6.10. Mimetic Interaction Spaces: Controlling Distant Displays in Pervasive Environments

Pervasive computing is a vision that has been an inspiring long-term target for many years now. Interaction techniques that allow one user to efficiently control many screens, or that allow several users to collaborate on one distant screen, are still hot topics, and are often considered as two different questions. Standard approaches require a strong coupling between the physical location of input device, and users. We propose to consider these two questions through the same basic concept, that uncouples physical location and user input, using a mid-air approach. We present the concept of mimetic interaction spaces (MIS), a dynamic user-definition of an imaginary input space thanks to an iconic gesture, that can be used to define mid-air interaction techniques. We describe a participative design user-study, that shows this technique has interesting acceptability and elicit some definition and deletion gestures. We finally describe a design space for MIS-based interaction, and show how such concept may be used for multi-screen control, as well as screen sharing in pervasive environments [26].

6.11. Match-Up & Conquer: A Two-Step Technique for Recognizing Unconstrained Bimanual and Multi-Finger Touch Input

We present a simple, two-step technique for recognizing multi-touch gesture input independently of how users articulate gestures, i.e., using one or two hands, one or multiple fingers, synchronous or asynchronous stroke input. To this end, and for the first time in the gesture literature, we introduce a preprocessing step specifically for multi-touch gestures (Match-Up) that clusters together similar strokes produced by different fingers, before running a gesture recognizer (Conquer). We report gains in recognition accuracy of up to 10% leveraged by our new preprocessing step, which manages to construct a more adequate representation for multi-touch gestures in terms of key strokes. It is our hope that the Match-Up technique will add to the practitioners toolkit of gesture preprocessing techniques, as a first step toward filling todays lack of algorithmic knowledge to process multi-touch input and leading toward the design of more efficient and accurate recognizers for touch surfaces. [27]

6.12. Understanding Users's perceived Difficulty of Multi-Touch Gesture Articulation

We show that users are consistent in their assessments of the articulation difficulty of multi-touch gestures, even under the many degrees of freedom afforded by multi-touch input, such as (1) various number of fingers touching the surface, (2) various number of strokes that structure the gesture shape, and (3) single-handed and bimanual input. To understand more about perceived difficulty, we characterize gesture articulations captured under these conditions with geometric and kinematic descriptors computed on a dataset of 7,200 samples of 30 distinct gesture types collected from 18 participants. We correlate the values of the objective descriptors with users' subjective assessments of articulation difficulty and report path length, production time, and gesture size as the highest cor- relators (max Pearson's r=.95). We also report new findings about multi-touch gesture input, e.g., gestures produced with more fingers are larger in size and take more time to produce than single-touch gestures; bimanual articulations are not only faster than single-handed input, but they are also longer in path length, present more strokes, and result in gesture shapes that are deformed horizontally by 35% in average. We use our findings to outline a number of 14 guidelines to assist multi-touch gesture set design, recognizer development, and inform gesture-to-function mappings through the prism of the user-perceived difficulty of gesture articulation.[28]

6.13. Dynamic Modelling of Electrovibration

Electrostatic attraction may be used to modulate the apparent friction coefficient between a fingertip and a surface to create a tactile stimulator. In this work, we want to propose an accurate modelling of the force generation. For that purpose, a specific experimental test bench has been manufactured, as presented in figure 5 .
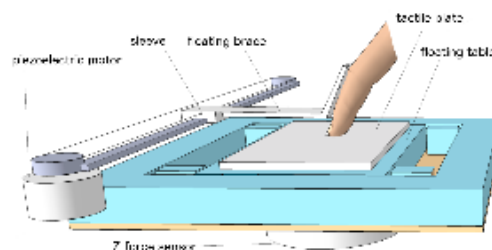

Figure 5. Representation of the measurement system, the finger is moved on the plate by the motor.

Then, an investigation on the current modeling were carried out, with a focus on the temporal evolution and frequency dependence of the stimulus. More particularly, we considered the charge lost through the stratum corneum. Indeed, lost charges is gathered on the surface of the insulator as free surface charge, for this reason it no longer participates to the generation of the force on the finger, and consequently, to the measured force (Fig. 6). This happens because the charges on the surface of the insulator are no longer mechanically bounded to the finger and the insulator sustains the induced electrostatic force.
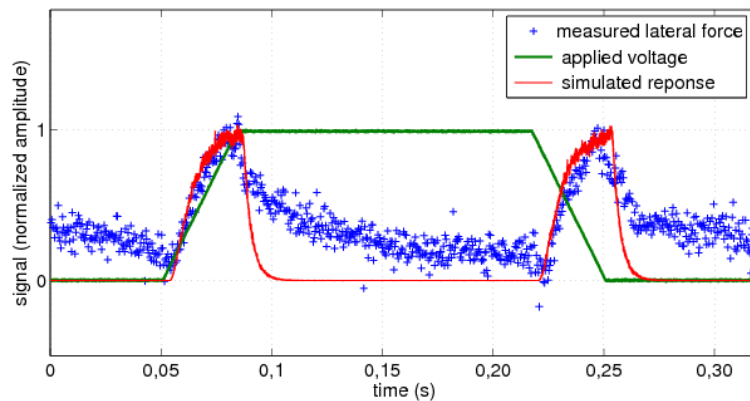

Figure 6. Charge configuration at the border of the stratum corneum (SC) and insulator (I). The conductive part of the system is represented like the electrode of a capacitor. (1) Initial configuration on the charge when the voltage v is applied. (2) Discharge through the stratum corneum with the two equivalent capacitors. (3) Final configuration of the charges after the transient.

The improvement of the modeling is proposed to take into account this major effect, and then, it is checked with an experimental set-up and compared with literature results.

6.14. Coupling between Electrovibration and squeeze filmfor tactile stimulation

Electrovibration and squeeze film effect are two different principles which modify user perception of a surface. The first is generated by a polarization of a finger approaching a high voltage supplied plate, and the latter by an ultrasonic vibrating plate. Their compatibility on the same stimulator has been analysed and their concomitant has been proven as well as the increased range of sensations [34]. A joint model has been proposed to describe the behaviour of the friction when both principles are merged. For the analysis, a specific experimental test bench has been built to measure the forces induced, as shown in figure 7 .
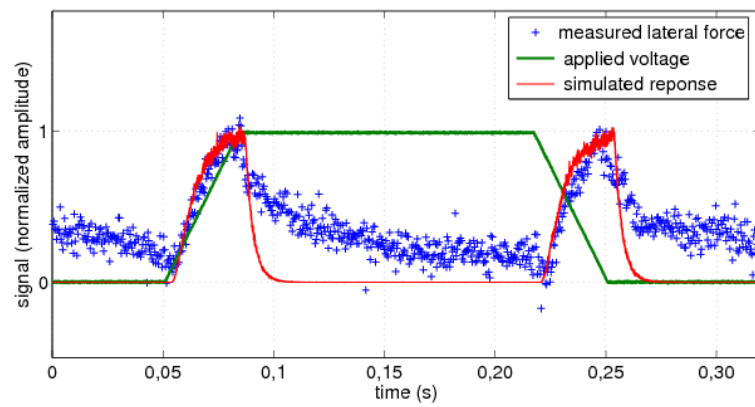
Figure 7. The experimental setup, and the recorded friction modulation.

MODAL Project-Team

6. New Results

6.1. Highlights of the Year

Thanks to the development technological action MPAGenomics, the team has created one of the first french instances of Galaxy publicly available on the French Bioinformatics cloud. This instance is original as it offers complex statistical tools for genomic data analysis in a user-friendly interface (see 5.9).

The team obtained bilateral contracts with companies as Auchan or RougeGorge thanks to its just emerging, but promising, clustering software MixtComp (see 5.14), dedicated to full mixed and missing data.

6.2. Model for conditionally correlated categorical data

Participants: Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle.

An extension of the latent class model is proposed for clustering categorical data by relaxing the classical class conditional independence assumption of variables. In this model (called CCM for Conditional Correlated Model), variables are grouped into inter-independent and intra-dependent blocks in order to consider the main intra-class correlations. The dependence between variables grouped into the same block is taken into account by mixing two extreme distributions, which are respectively the independence and the maximum dependence ones. In the conditionally correlated data case, this approach is expected to reduce biases involved by the latent class model and to produce a meaningful model with few additional parameters. The parameters estimation by maximum likelihood is performed by an EM algorithm while a MCMC algorithm avoiding combinatorial problems involved by the block structure search is used for model selection. Applications on sociological and biological data sets bring out the proposed model interest. These results strengthen the idea that the proposed model is meaningful and that biases induced by the conditional independence assumption of the latent class model are reduced. This work has been now accepted in an international journal [24]. Furthermore, an R package (Clustericat) is available on CRAN (see 5.3).

6.3. Model for conditionally correlated categorical data

Participants: Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle.

It is a model-based clustering proposal (called CMM for Conditional Modes Model) where categorical data are grouped into conditionally independent blocks. The corresponding block distribution is a parsimonious multinomial distribution where the few free parameters correspond to the most likely modality crossings, while the remaining probability mass is uniformly spread over the other modality crossings. The exact computation of the integrated complete-data likelihood allows to perform the model selection, by a Gibbs sampler, reducing the computing time consuming by parameter estimation and avoiding BIC criterion biases pointed out by our experiments. An article has been now submitted to an international journal [49]. Furthermore, an R package (CoModes) is available on Rforge (see 5.4).

6.4. Mixture model for mixed kind of data

Participants: Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle.

A mixture model of Gaussian copula allows to cluster mixed kind of data. Each component is composed by classical margins while the conditional dependencies between the variables is modeled by a Gaussian copula. The parameter estimation is performed by a Gibbs sampler. An article has been presented to an international conference [48] and has been also submitted to an international journal [50]. Furthermore, an R package (MixCluster) is available on Rforge (see 5.12).

6.5. Mixture of Gaussians with Missing Data

Participants: Christophe Biernacki, Vincent Vandewalle.

The generative models allow to handle missing data. This can be easily performed by using the EM algorithm, which has a closed form M-step in the Gaussian setting. This can for instance be useful for distance estimation with missing data. It has been proposed to improve the distance estimation by fitting a mixture of Gaussian distributions instead of a considering only one Gaussian component [16]. This is a joint work with Emil Eirola and Amaury Lendrasse .

6.6. Clustering and variable selection in regression

Participants: Christophe Biernacki, Loïc Yengo, Julien Jacques.

A new framework is proposed to address the issue of simultaneous linear regression and clustering of predictors where regression coefficients are assumed to be drawn from a Gaussian mixture distribution. Prediction is thus performed using the conditional distribution of the regression coefficients given the data, while clusters are easily derived from posterior distribution in groups given the data. This work is now published in [27]. Furthermore, an R package (clere) is available on Rforge (see 5.2) and an improved version of the initial model has been submitted to an international journal [52].

6.7. Model-based clustering for multivariate partial ranking data

Participants: Christophe Biernacki, Julien Jacques.

The first model-based clustering algorithm dedicated to multivariate partial ranking data is now published in an internation journal [19]. This is an extension of the (ISR) model for ranking data published in 2013. The proposed algorithm has allowed to exhibit regional alliances between European countries in the Eurovision contest, which are often suspected but never proved.

6.8. Generative models for correlated variables in regression

Participants: Christophe Biernacki, Clément Théry.

Linear regression outcomes (estimates, prevision) are known to be damaged by highly correlated covariates. However most modern datasets are expected to mechanically convey more and more highly correlated covariates due to the global increase of the amount of variables they contain. We propose to explicitly model such correlations by a family of linear regressions between the covariates. It leads to a particular generative model through the distribution explicitly introduced between correlated covariates. It has been presented to a conference [32] and is currently written as a research paper [51]. Furthermore, an R package (CorReg) is available on CRAN (see 5.5). Extension is now available for missing covariables also. It is a joint work with Gaétan Loridant.

6.9. Model-based clustering for multivariate partial ordinal data

Participants: Christophe Biernacki, Julien Jacques.

We design the first univariate probability distribution for ordinal data which strictly respects the ordinal nature of data. More precisely, it relies only on order comparisons between modalities, the proposed distribution being obtained by modeling the data generating process which is assumed, from optimality arguments, to be a stochastic binary search algorithm in a sorted table. The resulting distribution is natively governed by two meaningful parameters (position and precision) and has very appealing properties: decrease around the mode, shape tuning from uniformity to a Dirac, identifiability. Moreover, it is easily estimated by an EM algorithm since the path in the stochastic binary search algorithm is missing. Using then the classical latent class assumption, the previous univariate ordinal model is straightforwardly extended to model-based clustering for multivariate ordinal data. Again, parameters of this mixture model are estimated by an EM algorithm. Both simulated and real data sets illustrate the great potential of this model by its ability to parsimoniously identify particularly relevant clusters which were unsuspected by some traditional competitors. This work is currently in revision in an international journal [38].

6.10. Clustering for functional data into discriminative subspaces

Participant: Julien Jacques.

This is a joint work with Charles Bouveyron (Paris 5) and Etienne Côme (Inrets).

A model-based clustering method for time series has been developed, based on a discriminative functional mixture model which allows the clustering of the data in a functional subspace. This model presents the advantage to be parsimonious and can therefore handle long time series. This model has been used for analyzing different bike sharing systems In Europe.

6.11. Degeneracy in multivariate Gaussian mixtures

Participant: Christophe Biernacki.

In the case of Gaussian mixtures, unbounded likelihood is an important theoretical and practical problem. Using the weak information that the latent sample size of each component has to be greater than the space dimension, we derive a simple non-asymptotic stochastic lower bound on variances. We prove also that maximizing the likelihood under this data-driven constraint leads to consistent estimates. This work has been presented to a conference [31]. This is a joint work with Gwënaelle Castellan.

6.12. Auto-Associative Models

Participant: Serge Iovleff.

Auto-Associative models cover a large class of methods used in data analysis, among them are for example the famous PCA and the auto-associative neural networks. We describe the general properties of these models when the projection component is linear and we propose and test an easy to implement Probabilistic Semi-Linear Auto-Associative model in a Gaussian setting. This work is now published in [18].

6.13. Resampling and density estimation

Participant: Alain Celisse.

We characterized the behavior of cross-validation (Lpo) in density estimation with the L^2-loss. We considered two aspects: risk estimation and model selection. For the first one, we settled leave-one-out is optimal. On the contrary for the second one, we provided the first guidelines toward an optimal choice of the parameter p. In particular, this choice depends on the convergence rate of the best estimator in the family we consider.

6.14. Resampling and classification

Participant: Alain Celisse.

This is a joint work with Tristan Mary-Huard (INRA).

We extended known results about leave-one-out to the case of leave-p-out for the k-nearest neighbor estimator in classification with the 0-1 loss. In particular, our strategy relies on the relationship between leave-p-out and U-statistics. We derive upper bounds on the moments on the leave-p-out estimator as well as an exponential concentration inequality.

6.15. Kernel change-point

Participants: Alain Celisse, Guillemette Marot.

This is a joint work with Guillem Rigaill and Morgane Pierre-Jean (Univ. Evry).

Based on a previous work, we successfully applied kernel methods to change-point detection in the context of next generation sequencing with multivariate complex data. We also provided greatly improved algorithm in terms of computational complexity (both in time and space). With very huge amounts of data, we also suggest a new strategy based on the idea of approximating the Gram matrix by a low-rank matrix, which leads to a linear time complexity algorithm.

6.16. Normality test in RKHS

Participants: Alain Celisse, Jérémie Kellner.

In the kernel method framework, we use the MMD (maximum mean discrepancy) to derive a new goodness-of-fit test that can be used in the RKHS. When applied to the usual R^d setting, our test does not seem too sensitive to any increase on the dimension d unlike other ongoing approaches. With an infinite dimension RKHS, it exhibits a good power for a prescribed level of type-I error control.

6.17. Differential meta-analysis of RNA-seq data from multiple studies

Participant: Guillemette Marot.

This is a joint work with Andrea Rau and Florence Jaffrézic (INRA, Jouy-en-Josas).

An adaptation of meta-analysis methods initially proposed for microarray studies has been proposed for RNA-seq data. The research paper has been published in [26] and the associated R package metaRNASeq is now available on CRAN (see 5.11).

6.18. Multi-patient analysis of genomic markers

Participants: Quentin Grimonprez, Samuel Blanck, Guillemette Marot, Alain Celisse.

Tests performed during Development Technological Action MPAgenomics have shown on real data that it was also important to suggest automatic and appropriate calibrations for parameters in segmentation methods than to look for common markers able to predict patient's response. In the R package MPAgenomics (see 5.15), we have thus proposed two independent pipelines described in [17]. The choice of a given pipeline depends on the heterogeneity degree of studied genomic profiles.

6.19. Scan statistics for dependent data

Participants: Alexandru Amarioarei, Cristian Preda.

Dependent models of type block factors are introduced for scan statistics as an extension of the models based on the independent and identically distributed assumption. Approximations and errors are derived for one and two dimensions. Matlab software has been developed for this purpose.

NON-A Project-Team

6. New Results

6.1. Highlights of the Year

- We are becoming world-recognized on homogeneous approach to estimation and control [13], [24].
- New method of stability analysis and control design for time-delay systems: Implicit Lyapunov-Krasovski Functionals [72].
- New dynamical model of population of oysters for water quality monitoring [44].
- New local path planning algorithm for mobile robots based on intermediate objectives [33].
- New patent on method and device for detecting a failure on an aircraft [85].
- New book on robust control design [82].

6.2. Homogeneity Theory And Analysis Of Nonlinear Systems

Homogeneity is a kind of symmetry, if it is presented in a system model, then it may simplify analysis of stability and performance properties of the system. The new results obtained in 2014 are as follows:

- The problem of scalability of trajectories in homogeneous and locally homogeneous systems is considered [46]. It is shown that the homogeneous systems have scalability property, and locally homogeneous systems possess this property approximately.
- Constructive conditions for verification of input-to-state stability property for discontinuous systems using geometric homogeneity have been proposed in [48]. The characterization of the asymptotic gain for such systems has been presented in [47].
- The problem of finite-time output stabilization of the double integrator is addressed in [14] applying the homogeneity approach. Robustness and effects of discretization on the obtained closed loop system are analyzed.
- The paper [24] extends notion of homogeneity to the time-delay nonlinear systems. Generalizations and specifications of the homogeneity approach to time-delay nonlinear systems are given in [57], where, for instance, the stability independently on delay has been analyzed.
- In [75] the uniform stability notion for a class of non-linear time-varying systems is studied using the homogeneity framework. The results are applied to the problem of adaptive estimation for a linear system.
- The Implicit Lyapunov Function (ILF) method has been applied for homogeneous differentiator design [70]. The procedure for adjustment of differentiator parameters has been resented in the form of semi-definite programming problem. ILF-based algorithms of robust finite-time and fixed-time stabilization of the chain of integrators were developed in [34]. In [69] they were adapted for the second order sliding mode control design.
- The tutorial on homogeneous methods in high sliding mode control has been published [13]. It stresses some recently obtained results of the team about homogeneity for differential inclusions and robustness with respect to perturbations in the context of input-to-state stability.

6.3. Model-Free Control

The model free control techniques form a new and quickly developing area of control theory. It has been established by the team members and nowadays these tools find many practical applications and attract a lot of attention due to their clear advantages for designers: they provide a control law independently in the model knowledge. The achievements obtained in 2014 are as follows:

- The paper [67] proposes a motion planning approach for non-holonomic mobile robots using i-PID controller. The effectiveness and the robustness of the proposed method are shown via several simulations.

- In [60] we show that the open-loop transfer functions and the stability margins may be defined within the recent model-free control setting. Several convincing computer experiments are presented including one which studies the robustness with respect to delays.

6.4. Algebraic Technique For Estimation, Differentiation And Its Applications

Elementary techniques from operational calculus, differential algebra, and non-commutative algebra lead to a new algebraic approach for estimation and detection. It is investigated in various areas of applied sciences and engineering. The following lists only some applications:

- The article [19] presents an algebraic on-line parameters estimation method for Linear Time Invariant (LTI) systems subject to polynomial perturbations. Particular attention is given to practical implementation.

- In the paper [43], we extend the modulating functions method to estimate the state and the unknown input of a linear time-varying system defined by a linear differential equation. Numerical results are given to show the accuracy and the robustness of the proposed estimators against corrupting noises.

- The paper [17] presents a new sensorless parameter identification method for permanent magnet stepper motors. The method exploits derived linear parameterizations and least-squares algorithms.

- In [36] a continuous-time least-squares parameter estimation method through evolution equations is proposed. A deterministic framework for the estimation under noisy measurements is proposed using a Sobolev space with negative index to model of the noise.

- Causation between time series is a most important topic in econometrics, financial engineering, biological and psychological sciences, and many other fields. A new setting is introduced in [42] for examining this rather abstract concept. The corresponding calculations, which are much easier than those required by the celebrated Granger-causality, do not necessitate any deterministic or probabilistic modeling.

- The paper [59] proposes a solution to the problem of velocity and position estimation for a class of oscillating systems whose position, velocity and acceleration are zero mean signals. The proposed scheme considers that the dynamic model of the system is unknown and only noisy acceleration measurements are available.

- The communications [63], [78] are devoted to solar irradiance and irradiation short-term forecasts, which are useful for electricity production. Several different time series approaches are employed.

- In [68] we present a simple algorithm to compute the factors of a Unimodular-Upper polynomial matrix decomposition. Such decomposition is useful for spatial multiplexing in multi-input multi-output(MIMO) channel transmission system since it enables to reduce the MIMO channel matrix into independent channels by a pre- and post-filtering.

- A fault-tolerant control method based on algebraic derivative estimation is introduced in [32]. It is applied on an electromagnetically supported plate as an example of a nonlinear and an open-loop unstable system.

6.5. Observability And Observer Design For Nonlinear Systems

Observability analysis and observer design are important issues in the field of control theory. Some recent results are listed below:

- The paper [12] deals with the observability analysis of linear time systems whose outputs are affected by unknown inputs. Three different definitions of observability are proposed. Sufficient conditions are deduced for each proposed definition.

- In [11] a method of the state estimation is proposed for a class of nonlinear systems with unknown inputs whose dynamics is governed by differential-algebraic equations (DAE). The estimation is done using a sliding mode high order differentiator.

- The recent algebraic parametric method proposed by Fliess and Sira-Ramirez has been extended to numerical differentiation problem in noisy environment [66]. The obtained algebraic differentiators are non-asymptotic and robust against corrupting noises.

- The paper [41] investigates the observer design problem of nonlinear impulsive systems with impact perturbation. By using the concept of normal form, it proposes a full order finite time observer, which guarantees the finite time convergence independent of the impact perturbation.

- The development of adaptive observer techniques for nonlinear systems in the output canonical form is proposed in [22] applying additional impulsive feedback in the observer equations. The stability is investigated.

- In [55] the problem of adaptive observer design in the presence of disturbances is studied, and an augmented adaptive observer is proposed using sliding mode methodology.

6.6. Sliding Mode Control And Estimation

Sliding mode algorithms are very popular for finite-time estimation and regulation. The recent results obtained by the group are as follows:

- In [71] the high-order sliding mode control design algorithm has been developed for MIMO system using ILF Method. Procedure for tuning of control parameters is presented using Linear Matrix Inequalities.

- A novel hybrid automaton admitting the modeling of both conventional and modern(high order) sliding mode systems is presented [65]. A scheme for defining hybrid-automaton executions beyond Zeno points is proposed by means of introduction of Filippov-like executions.

- The paper [35] surveys mathematical tools required for stability/convergence analysis of modern sliding mode control systems and introduces the generalized Lyapunov theorems. Application of these results to finite-time stability analysis and settling time estimation of twisting second order sliding mode controller are given [73].

- The problem of the sliding mode control design is considered in [81] for the linear time-invariant disturbed system with the noised measurements of the output. The control law, which provides to the closed-loop system the optimal reaching (as close as possible) of the selected sliding surface, is designed using minimax state observer.

- The paper [50] deals with a signal-based method for robust and early detection of lock-in-place failures (a.k.a. jamming) in aircraft control surface servo-loops. The signal-based scheme is proposed using a sliding-mode differentiator. The developed monitoring scheme has been tested on Airbus test facilities located at Toulouse, France.

- In the paper [79], we investigate the problem of adaptive observer for simultaneous estimation of state and parameter for a class of nonlinear systems.Necessary condition for the existence of such an observer is derived. The paper [76] uses developed technique for states estimation and parameter identification for nonlinear Dengue epidemic model.

- The paper [80] investigates the problem of global finite-time observer design for a class of nonlinear systems which can be transformed into the output depending normal form.

6.7. Non-Linear, Sampled And Time-Delay Systems

Nonlinearities, sampling, quantization and time-delays cause serious obstructions for control and observer design in many fields of techniques and engineering (e.g. networked and internet systems, distributed systems etc.). The proposed by the team algebraic approach suits well for estimation and regulation in such a type of systems. The recent results are listed below:

- The method of Implicit Lyapunov-Krasovski Functional for stability analysis of time-delay systems is introduced in [72].

- The article [31] proposes a convex optimization approach for the design of relay feedback controllers. Furthermore, the approach is used in the sampled-data case in order to guarantee (locally) the practical stabilization to a bounded ellipsoid of the order of the sampling interval.

- The paper [40] addresses the controller design problem for bilateral teleoperation over unreliable networks. The stability and tracking performance analysis are presented for a novel force-reflecting emulator control scheme.

- The problem of time optimal control design is considered for a chain of integrators in [74]. The suboptimal continuous ILF-based solution is presented and compared with the optimal discontinuous feedback.

- In the erratum [26] recently proposed conditions on finite-time stability in time-delay systems are revisited and it is shown that they are incorrect. General comments on possibility of finite-time convergence in time-delay systems and a necessary condition are given.

- The problem of formulation of an equivalent characterization for instability is considered in [56]. The necessary part of the Chetaev's theorem on instability is formulated. Using the developed necessary instability conditions, the Anti-control Lyapunov Function (ALF) framework is extended and the Control Chetaev Function (CCF) concept is proposed as a counterpart of the Control Lyapunov function (CLF) theory.

- The paper [25] extends the notion of oscillations in the sense of Yakubovich to hybrid dynamics. Several sufficient stability and instability conditions for a forward invariant set are presented. The consideration is illustrated by analysis of a model of two-link compass-gait biped robot.

- The paper [15] deals with the design of an active fault-tolerant control strategy based on the supervisory control approach technique for linear time invariant MIMO systems affected by disturbances, measurement noise, and faults.

- The problem of phase regulation for a population of oscillating systems is considered in [21]. The proposed control strategy is based on a Phase Response Curve (PRC) model of an oscillator.

- The paper [51] deals with the design of an estimator-based supervisory Fault Tolerant Control scheme for Linear Time Invariant systems. A formal stability proof based on dwell-time conditions is established.

- In [39], we propose a general statistical framework for model based compressive sensing, where both sparsity and structure priors are considered simultaneously. It is based on the Latent Variable Analysis and the Gamma-Gaussian modelling.

- The paper [64] investigates the left invertibility for nonlinear time delay system with internal dynamics under some assumptions imposed on the internal dynamics. Causal and non causal estimation of the unknown inputs are respectively discussed, and the high-order sliding mode observer is used to estimate the observable states.

- In the paper [54] a simple second order model is proposed for modeling the pressure dynamics with a pure time delay on the control input. The Artstein transformation is applied in order to design the stabilizing robust nonlinear controller.

6.8. Set-Theoretic Methods of Control And Estimation

In many cases due to parametric and/or signal uncertainties presented in a plant model it is not possible to design a conventional observer, which provides a point-wise estimate of state in a finite time or asymptotically. In this case it is still frequently possible to design observers, which generate an estimate on the set of the admissible values of the state at the current instant of time. The recent new results in this field are listed below:

- An interval observer for Linear Time-Varying systems is proposed in [38]. A constructive approach to obtain a time-varying change of coordinates, ensuring the cooperativity of the observer error in the new coordinates, is provided in order to simplify the design of the interval observer.

- In [58] the problem of interval observer design is addressed for a class of descriptor linear systems with delays. An interval observation for any input in the system is provided. The control input is designed together with the observer gains in order to guarantee interval estimation and stabilization simultaneously.

- The estimation problem of a system with unknown time-delay and unknown input gains is considered in [49]. The interval observation technique is applied in order to obtain guaranteed interval of the system state.

- The book [82] introduces newly developed robust control design technique for a wide class of continuous-time dynamical systems called the "attractive ellipsoid method." It studies nonlinear affine control systems in the presence of uncertainty and presents a constructive and easily implementable control strategy that guarantees certain stability properties.

6.9. Networked Robots

The mobile robots constitute an important area of practical development for the team:

- The paper [33] presents a path planning algorithm for autonomous navigation of non-holonomic mobile robots in complex environments. The irregular contour of obstacles is represented by segments. The optimal path planning problem is formulated as a constrained receding horizon planning problem and the trajectory is obtained by solving an optimal control problem with constraints.

- In [62] robot dynamic parameters are estimated based on power model associated with modulating functions, which avoids measuring or calculating the joint acceleration. At the same time, an advanced causal Jacobi derivative estimator is applied in order to get on-line robust derivatives from noisy measurements.

- The paper [61] provides a solution for the stabilization of a nonholonomic wheeled mobile robot which is affected by additive input disturbances. The solution is based on the supervisory control framework, finite-time stability and robust multi-output regulation.

- The demo video with the developments of NON-A team in networked robotics is given by https://www.youtube.com/watch?v=Mq_hB0UkzkY

6.10. Applications

As it was mentioned, Non-A is a kind of "method-driven" project, which deals with different aspects of finite-time estimation and control. Thus different applications are possible, ones touched this year are as follows (skipping the networked robots considered in the previous section):

- Method and device for detecting a failure on an aircraft are developed and patented [85].

- In [44] the measurements of valve activity in a population of bivalves under natural environmental conditions (16 oysters in the Bay of Arcachon, France) are used for a physiological model identification. A nonlinear auto-regressive exogenous (NARX) model is designed and tested. The developed dynamical model can be used for estimation of the normal physiological rhythms of permanently immersed oysters and, in particular, for ecological monitoring.

- The articles [53], [18], [20], [77] present novel control strategies for Permanent Magnet Synchronous Motor (PMSM), which does not ignores the relay nature of the actuators. A design procedure based on Linear Matrix Inequalities (LMI) allows us to derive the switching surfaces, which depend on the motor position. The sliding mode and nonlinear adaptive observers are designed for state estimation and parameters identification.

- The problem of air-fuel ratio stabilization in spark ignition engines is addressed in the paper [23]. The proposed strategy consists of proper switching among two control laws. The first one is based on an a priori off-line identified engine model and the second control law is adaptive. The supervisor realizes a switching rule between them providing better performance. Results of implementation on two vehicles are reported and discussed.

- The paper [37] deals with a control design for serial multicellular choppers. The novel scheme that uses two Petri nets (PNs) to carry out the control action is introduced. Experimental results from four and five-level choppers are used to emphasize the performance and the effectiveness of the proposed control scheme.

- The paper [52] is concerned with preliminary results on robot vibratory modes on-line identification using the external measurement provided by a laser tracker. A comparison between the algebraic method and the sliding modes for the parameter identification is proposed. Experimental identifications are proposed on a 6 degrees of freedom (DOF) manipulator robot Stäubli RX-170B.

- The papers [30], [29], [16] develop different fault detection schemes for robust and early detection of faults in aircraft control surfaces servo-loop. A complete Monte Carlo campaign from a high representative simulator, provided by Airbus as a part of the ADDSAFE project, as well as experimental results obtained on AIRBUS test facilities demonstrate the high fault detection performance, robustness and viability of the proposed techniques.

- The paper [28] deals with the problem of the practical tracking control of an experimental car-like system called the Robucar - a four-wheeled car in a single steering mode. A practical tracking controller is designed using the second-order sliding mode control. Experimental tests are presented and compared with the conventional sliding controller.

- Power converters are a very important for the control of high power systems. In the article [45] we propose a control strategy for minimizing the no-load conduction losses and analyze the transient behavior in case of load steps including output short-circuit.

<p style="text-align:center; color:red;">RMOD Project-Team</p>

6. New Results

6.1. Highlights of the Year

- Pharo 3.0 has been released in April 2014.
- Moose 5.0 has been released in December 2014.
- The book Deep into Pharo has been released publicly http://www.deepintopharo.com.
- RMOD entered in a sponsoring agreement with LAM Research, Inc.

6.2. Tools for understanding applications

Remodularization Analysis Using Semantic Clustering. We report an experience on using and adapting Semantic Clustering to evaluate software remodularizations. Semantic Clustering is an approach that relies on information retrieval and clustering techniques to extract sets of similar classes in a system, according to their vocabularies. We adapted Semantic Clustering to support remodularization analysis. We evaluate our adaptation using six real-world remodularizations of four software systems. We report that Semantic Clustering and conceptual metrics can be used to express and explain the intention of the architects when performing common modularization operators, such as module decomposition. [37]

Towards a new package dependency model. Smalltalk originally did not have a package manager. Each Smalltalk implementation defined its own with more or less functionalities. Since 2010, Monticello/Metacello[Hen09] one package manager is available for open-source Smalltalks. It allows one to load source code packages with their dependencies. This package manager does not have all features we can find in well-known package managers like those used for the Linux operating system. We identify the missing features and propose a solution to reach a full-featured package manager. A part of this solution is to represent packages and dependencies as first-class objects, leading to the definition of a new dependency model. [32]

A Domain Specific Aspect Language for IDE Events. Integrated development environments (IDEs) have become the primary way to develop software. Besides just using the built-in features, it becomes more and more important to be able to extend the IDE with new features and extensions. Plugin architectures exist, but they show weaknesses related to unanticipated extensions and event handling. We argue that a more general solution for extending IDEs is needed. We present and discuss a solution, motivated by a set of concrete examples: a domain specific aspect language for IDE events. In it, join points are events of interest that may trigger the advice in which the behavior of the IDE extension is called. We show how this allows for the development of IDE plugins and demonstrate the advantages over traditional publish/subscribe systems. [21]

AspectMaps: Extending Moose to visualize AOP software. When using aspect-oriented programming the application implicitly invokes the functionality contained in the aspects. Consequently program comprehension of such a software is more intricate. To alleviate this difficulty we developed the AspectMaps visualization and tool. AspectMaps extends the Moose program comprehension and reverse engineering platform with support for aspects, and is implemented using facilities provided by Moose. We present the AspectMaps tool, and show how it can be used by performing an exploration of a fairly large aspect-oriented application. We then show how we extended the FAMIX meta-model family that underpins Moose to also provide support for aspects. This extension is called ASPIX, and thanks to this enhancement Moose can now also treat aspect-oriented software. Finally, we report on our experiences using some of the tools in Moose; Mondrian to implement the visualization, and Glamour to build the user interface. We discuss how we were able to implement a sizable visualization tool using them and how we were able to deal with some of their limitations. [20]

6.3. Software Quality: Taming Software Evolution

APIEvolutionMiner: Keeping API Evolution under Control. During software evolution, source code is constantly refactored. In real-world migrations, many methods in the newer version are not present in the old version (e.g., 60% of the methods in Eclipse 2.0 were not in version 1.0). This requires changes to be consistently applied to reflect the new API and avoid further maintenance problems. We propose a tool to extract rules by monitoring API changes applied in source code during system evolution. In this process, changes are mined at revision level in code history. Our tool focuses on mining invocation changes to keep track of how they are evolving. We also provide three case studies in order to evaluate the tool. [34]

Towards an Automation of the Mutation Analysis Dedicated to Model Transformation. A major benefit of Model Driven Engineering (MDE) relies on the automatic generation of artefacts from high-level models through intermediary levels using model transformations. In such a process, the input must be well-designed and the model transformations should be trustworthy. Due to the specificities of models and transformations, classical software test techniques have to be adapted. Among these techniques, mutation analysis has been ported and a set of mutation operators has been defined. However, mutation analysis currently requires a considerable manual work and suffers from the test data set improvement activity. This activity is seen by testers as a difficult and time-consuming job, and reduces the benefits of the mutation analysis. This work addresses the test data set improvement activity. Model transformation traceability in conjunction with a model of mutation operators, and a dedicated algorithm allow to automatically or semi-automatically produce test models that detect new faults. The proposed approach is validated and illustrated in a case study written in Kermeta. [17]

Predicting software defects with causality tests. We propose a defect prediction approach centered on more robust evidences towards causality between source code metrics (as predictors) and the occurrence of defects. More specifically, we rely on the Granger causality test to evaluate whether past variations in source code metrics values can be used to forecast changes in time series of defects. Our approach triggers alarms when changes made to the source code of a target system have a high chance of producing defects. We evaluated our approach in several life stages of four Java-based systems. We reached an average precision greater than 50% in three out of the four systems we evaluated. Moreover, by comparing our approach with baselines that are not based on causality tests, it achieved a better precision. [19]

6.4. Software Quality: History and Changes

Tracking dependencies between code changes: An incremental approach. Merging a change often leads to the question of knowing what are the dependencies to other changes that should be merged too to obtain a working system. This question also arises with code history trackers – Code history trackers are tools that react to what the developer do by creating first-class objects that represent the change made to the system. We evaluate the capacity of different code history trackers to represent, also as first-class objects, the dependencies between those changes. We also present a representation for dependencies that works with the event model of Epicea, a fine-grained and incremental code history tracker. [32]

Mining Architectural Violations from Version History. Software architecture conformance is a key software quality control activity that aims to reveal the progressive gap normally observed between concrete and planned software architectures. However, formally specifying an architecture can be difficult, as it must be done by an expert of the system having a high level understanding of it. We present a lightweighted approach for architecture conformance based on a combination of static and historical source code analysis. The proposed approach relies on four heuristics for detecting absences (something expected was not found) and divergences (something prohibited was found) in source code based architectures. We also present an architecture conformance process based on the proposed approach. We followed this process to evaluate the architecture of two industrial-strength information systems, achieving an overall precision of 62.7% and 53.8%. We also evaluated our approach in an open-source information retrieval library, achieving an overall precision of 59.2%. We envision that aheuristic-based approach for architecture conformance can be used to rapidly raise architectural warnings, without deeply involving experts in the process. [22]

6.5. Reconciling Dynamic Languages and Isolation

Delegation Proxies: The Power of Propagation. Scoping behavioral variations to dynamic extents is useful to support non-functional requirements that otherwise result in cross-cutting code. Unfortunately, such variations are difficult to achieve with traditional reflection or aspects. We show that with a modification of dynamic proxies, called delegation proxies, it becomes possible to reflectively implement variations that propagate to all objects accessed in the dynamic extent of a message send. We demonstrate our approach with examples of variations scoped to dynamic extents that help simplify code related to safety, reliability, and monitoring. [38]

Reifying the Reflectogram. Reflective facilities in OO languages are used both for implementing language extensions (such as AOP frameworks) and for supporting new programming tools and methodologies (such as object-centric debugging and message-based profiling). Yet controlling the run-time behavior of these reflective facilities introduces several challenges, such as computational overhead, the possibility of meta-recursion and an unclean separation of concerns between base and meta-level. We present five dimensions of meta-level control from related literature that try to remedy these problems. These dimensions are namely: temporal and spatial control, placement control, level control and identity control. We argue that the reification of the descriptive notion of the reflectogram, can unify the control of meta-level execution in all these five dimensions. We present a model for the reification of the reflectogram and validate our approach through a prototype implementation in the Pharo programming environment. Finally we detail a case study on run-time tracing illustrating our approach. [35]

Bootstrapping Reflective Systems: The Case of Pharo. Bootstrapping is a technique commonly known by its usage in language definition by the introduction of a compiler written in the same language it compiles. This process is important to understand and modify the definition of a given language using the same language, taking benefit of the abstractions and expression power it provides. A bootstrap, then, supports the evolution of a language. However, the infrastructure of reflective systems like Smalltalk includes, in addition to a compiler, an environment with several self-references. A reflective system bootstrap should consider all its infrastructural components. We propose a definition of bootstrap for object-oriented reflective systems, we describe the architecture and components it should contain and we analyze the challenges it has to overcome. Finally, we present a reference bootstrap process for a reflective system and Hazelnut, its implementation for bootstrapping the Pharo Smalltalk-inspired system. [26]

6.6. Dynamic Languages: Virtual Machines

Benzo: Reflective Glue for Low-level Programming. The goal of high-level low-level programming is to bring the abstraction capabilities of high-level languages to the system programming domain, such as virtual machines (VMs) and language runtimes. However, existing solutions are bound to compilation time and expose limited possibilities to be changed at runtime and from language-side. They do not fit well with fully reflective languages and environments. We propose Benzo1, a lightweight framework for high-level low-level programming that allows developers to generate and execute at runtime low-level code. It promotes the implementation, and dynamic modification, of system components with high-level language tools outperforming existing dynamic solutions. Since Benzo is a general framework we choose three applications that cover an important range of the spectrum of system programming for validating the infrastructure: a Foreign Function Interface (FFI), primitives instrumentation and a just-in-time bytecode compiler (JIT). With Benzo we show that these typical VM-level components are feasible as reflective language-side implementations. Due to its unique combination of high-level reflection and low-level programming, Benzo shows better performance for these three applications than the comparable high-level implementations. [30]

A bytecode set for adaptive optimizations. The Cog virtual machine features a bytecode interpreter and a baseline Just-in-time compiler. To reach the performance level of industrial quality virtual machines such as Java HotSpot, it needs to employ an adaptive inlining com-piler, a tool that on the fly aggressively optimizes frequently executed portions of code. We decided to implement such a tool as a bytecode to bytecode optimizer, implemented above the virtual machine, where it can be written and developed in Smalltalk. The optimizer we plan needs to extend the operations encoded in the bytecode set and its quality heavily depends on the bytecode

set quality. The current bytecode set understood by the virtual machine is old and lacks any room to add new operations. We decided to implement a new bytecode set, which includes additional bytecodes that allow the Just-in-time compiler to generate less generic, and hence simpler and faster code sequences for frequently executed primitives. The new bytecode set includes traps for validating speculative inlining de-cisions and is extensible without compromising optimization opportunities. In addition, we took advantage of this work to solve limitations of the current bytecode set such as the maximum number of instance variable per class, or number of literals per method. We plan to have it in production in the Cog virtual machine and its Pharo, Squeak and Newspeak clients in the coming year. [43]

6.7. Traits

Trait-oriented Programming in Java 8 Java 8 was released recently. Along with lambda expressions, a new language construct is introduced: default methods in interfaces. The intent of this feature is to allow interfaces to be extended over time preserving backward compatibility. We show a possible, different use of interfaces with default methods: we introduce a trait-oriented programming style based on an interface-as-trait idea, with the aim of improving code modularity. Starting from the most common operators on traits, we introduce some programming patterns mimicking such operators and discuss this approach. [29]

6.8. Tailoring Applications

In the context of the PhD of G. Polito, we developed Tornado, a way to generate specialized and minimal runtime. Using a run-fail-grow approach, which tries to execute an expression in an empty world, and on failure copies the missing program elements from a mother environment to the currently empty world, we could grew 11k full reflective application adding two numbers or 18k for the 100 factorial expression. We also usd this approach to generate specialized webserver in around 500kb. These results show that we can generate hyperspecialized kernels.

<p style="text-align:center">**SEQUEL Project-Team**</p>

6. New Results

6.1. Highlights of the Year

- New startup by Rémi Coulom on AI in games (go, chess, mahjong, $\cdots$).
- Successful Collaboration with Deezer and the victory at the ACM RecSys Recommendation Systems Challenge
- We were selected and working on preparation of ICML 2015 in Lille. ICML is the most important conference in the field of machine learning. This is the first time after more than 30 years of existence, that this conference will be held in France.

6.2. Decision-making Under Uncertainty

6.2.1. Reinforcement Learning

Selecting Near-Optimal Approximate State Representations in Reinforcement Learning [23]

We consider a reinforcement learning setting where the learner does not have explicit access to the states of the underlying Markov decision process (MDP). Instead, she has access to several models that map histories of past interactions to states. Here we improve over known regret bounds in this setting, and more importantly generalize to the case where the models given to the learner do not contain a true model resulting in an MDP representation but only approximations of it. We also give improved error bounds for state aggregation.

Online Stochastic Optimization under Correlated Bandit Feedback [15]

In this paper we consider the problem of online stochastic optimization of a locally smooth function under bandit feedback. We introduce the high-confidence tree (HCT) algorithm, a novel anytime X -armed bandit algorithm, and derive regret bounds matching the performance of state-of-the-art algorithms in terms of the dependency on number of steps and the near-optimality di-mension. The main advantage of HCT is that it handles the challenging case of correlated bandit feedback (reward), whereas existing meth-ods require rewards to be conditionally indepen-dent. HCT also improves on the state-of-the-art in terms of the memory requirement, as well as requiring a weaker smoothness assumption on the mean-reward function in comparison with the existing anytime algorithms. Finally, we discuss how HCT can be applied to the problem of policy search in reinforcement learning and we report preliminary empirical results.

Sparse Multi-task Reinforcement Learning [9]

In multi-task reinforcement learning (MTRL), the objective is to simultaneously learn multiple tasks and exploit their similarity to improve the performance w.r.t. single-task learning. In this paper we investigate the case when all the tasks can be accurately represented in a linear approximation space using the same small subset of the original (large) set of features. This is equivalent to assuming that the weight vectors of the task value functions are *jointly sparse*, i.e., the set of their non-zero components is small and it is shared across tasks. Building on existing results in multi-task regression, we develop two multi-task extensions of the fitted Q-iteration algorithm. While the first algorithm assumes that the tasks are jointly sparse in the given representation, the second one learns a transformation of the features in the attempt of finding a more sparse representation. For both algorithms we provide a sample complexity analysis and numerical simulations.

6.2.2. Multi-arm Bandit Theory

Spectral Bandits for Smooth Graph Functions with Applications in Recommender Systems [20]

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this paper, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each recommended item is a node and its expected rating is similar to its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in real-world graphs, and propose two algorithms for solving our problem that scale linearly in this dimension. Our experiments on real-world content recommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens nodes evaluations.

Online combinatorial optimization with stochastic decision sets and adversarial losses [21]

Most work on sequential learning assumes a fixed set of actions that are available all the time. However, in practice, actions can consist of picking subsets of readings from sensors that may break from time to time, road segments that can be blocked or goods that are out of stock. In this paper we study learning algorithms that are able to deal with stochastic availability of such unreliable composite actions. We propose and analyze algorithms based on the Follow-The-Perturbed-Leader prediction method for several learning settings differing in the feedback provided to the learner. Our algorithms rely on a novel loss estimation technique that we call Counting Asleep Times. We deliver regret bounds for our algorithms for the previously studied full information and (semi-)bandit settings, as well as a natural middle point between the two that we call the restricted information setting. A special consequence of our results is a significant improvement of the best known performance guarantees achieved by an efficient algorithm for the sleeping bandit problem with stochastic availability. Finally, we evaluate our algorithms empirically and show their improvement over the known approaches.

Extreme bandits [10]

In many areas of medicine, security, and life sciences, we want to allocate limited resources to different sources in order to detect extreme values. In this paper, we study an efficient way to allocate these resources sequentially under limited feedback. While sequential design of experiments is well studied in bandit theory, the most commonly optimized property is the regret with respect to the maximum mean reward. However, in other problems such as network intrusion detection, we are interested in detecting the most extreme value output by the sources. Therefore, in our work we study extreme regret which measures the efficiency of an algorithm compared to the oracle policy selecting the source with the heaviest tail. We propose the ExtremeHunter algorithm, provide its analysis, and evaluate it empirically on synthetic and real-world experiments.

Efficient learning by implicit exploration in bandit problems with side observations [18]

We consider online learning problems under a a partial observability model capturing situations where the information conveyed to the learner is between full information and bandit feedback. In the simplest variant, we assume that in addition to its own loss, the learner also gets to observe losses of some other actions. The revealed losses depend on the learner's action and a directed observation system chosen by the environment. For this setting, we propose the first algorithm that enjoys near-optimal regret guarantees without having to know the observation system before selecting its actions. Along similar lines, we also define a new partial information setting that models online combinatorial optimization problems where the feedback received by the learner is between semi-bandit and full feedback. As the predictions of our first algorithm cannot be always computed efficiently in this setting, we propose another algorithm with similar properties and with the benefit of always being computationally efficient, at the price of a slightly more complicated tuning mechanism. Both algorithms rely on a novel exploration strategy called implicit exploration, which is shown to be more efficient both computationally and information-theoretically than previously studied exploration strategies for the problem.

Best-Arm Identification in Linear Bandits [29]

We study the best-arm identification problem in linear bandit, where the rewards of the arms depend linearly on an unknown parameter θ^* and the objective is to return the arm with the largest reward. We characterize the complexity of the problem and introduce sample allocation strategies that pull arms to identify the best arm with a fixed confidence, while minimizing the sample budget. In particular, we show the importance of exploiting the global linear structure to improve the estimate of the reward of near-optimal arms. We analyze the proposed strategies and compare their empirical performance. Finally, we point out the connection to the G-optimality criterion used in optimal experimental design.

Exploiting easy data in online optimization [28]

We consider the problem of online optimization, where a learner chooses a decision from a given decision set and suffers some loss associated with the decision and the state of the environment. The learner's objective is to minimize its cumulative regret against the best fixed decision in hindsight. Over the past few decades numerous variants have been considered, with many algorithms designed to achieve sub-linear regret in the worst case. However, this level of robustness comes at a cost. Proposed algorithms are often over-conservative, failing to adapt to the actual complexity of the loss sequence which is often far from the worst case. In this paper we introduce a general algorithm that, provided with a "safe" learning algorithm and an opportunistic "benchmark", can effectively combine good worst-case guarantees with much improved performance on "easy" data. We derive general theoretical bounds on the regret of the proposed algorithm and discuss its implementation in a wide range of applications, notably in the problem of learning with shifting experts (a recent COLT open problem). Finally, we provide numerical simulations in the setting of prediction with expert advice with comparisons to the state of the art.

Spectral Bandits for Smooth Graph Functions [32]

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this paper, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each item we can recommend is a node and its expected rating is similar to its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret with respect to the optimal policy would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in real-world graphs, and propose two algorithms for solving our problem that scale linearly and sublinearly in this dimension. Our experiments on real-world content recommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens of nodes evaluations.

Regret bounds for restless Markov bandits [5]

We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm, that first represents the setting as an MDP which exhibits some special structural properties. In order to grasp this information we introduce the notion of ϵ-structured MDPs, which are a generalization of concepts like (approximate) state aggregation and MDP homomorphisms. We propose a general algorithm for learning ϵ-structured MDPs and show regret bounds that demonstrate that additional structural information enhances learning. Applied to the restless bandit setting, this algorithm achieves after any T steps regret of order $\widetilde{O}(T^{1/2})$ with respect to the best policy that knows the distributions of all arms. We make no assumptions on the Markov chains underlying each arm except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

Spectral Thompson Sampling [19]

Thompson Sampling (TS) has surged a lot of interest due to its good empirical performance, in particular in the computational advertising. Though successful, the tools for its performance analysis appeared only recently. In this paper, we describe and analyze SpectralTS algorithm for a bandit problem, where the payoffs of the choices are smooth given an underlying graph. In this setting, each choice is a node of a graph and the expected payoffs of the neighboring nodes are assumed to be similar. Although the setting has application

both in recommender systems and advertising, the traditional algorithms would scale poorly with the number of choices. For that purpose we consider an effective dimension d, which is small in real-world graphs. We deliver the analysis showing that the regret of SpectralTS scales as $d(T \ln N)^{1/2}$ with high probability, where T is the time horizon and N is the number of choices. Since a d sqrt(T ln N) regret is comparable to the known results, SpectralTS offers a computationally more efficient alternative. We also show that our algorithm is competitive on both synthetic and real-world data.

6.2.3. Recommendation systems

User Engagement as Evaluation: a Ranking or a Regression Problem? [39]

In this paper, we describe the winning approach used on the RecSys Challenge 2014 which focuses on employing user en-gagement as evaluation of recommendations. On one hand, we regard the challenge as a ranking problem and apply the LambdaMART algorithm, which is a listwise model special-ized in a Learning To Rank approach. On the other hand, after noticing some specific characteristics of this challenge, we also consider it as a regression problem and use pointwise regression models such as Random Forests. We compare how these different methods can be modified or combined to improve the accuracy and robustness of our model and we draw the advantages or disadvantages of each approach.

Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques [22]

In many recommendation applications such as news recommendation, the items that can be recommended come and go at a very fast pace. This is a challenge for recommender systems (RS) to face this setting. Online learning algorithms seem to be the most straight forward solution. The contextual bandit framework was introduced for that very purpose. In general the evaluation of a RS is a critical issue. Live evaluation is often avoided due to the potential loss of revenue, hence the need for offline evaluation methods. Two options are available. Model based meth- ods are biased by nature and are thus difficult to trust when used alone. Data driven methods are therefore what we consider here. Evaluat- ing online learning algorithms with past data is not simple but some methods exist in the litera- ture. Nonetheless their accuracy is not satisfac- tory mainly due to their mechanism of data re- jection that only allow the exploitation of a small fraction of the data. We precisely address this issue in this paper. After highlighting the limita- tions of the previous methods, we present a new method, based on bootstrapping techniques. This new method comes with two important improve- ments: it is much more accurate and it provides a measure of quality of its estimation. The latter is a highly desirable property in order to minimize the risks entailed by putting online a RS for the first time. We provide both theoretical and ex- perimental proofs of its superiority compared to state-of-the-art methods, as well as an analysis of the convergence of the measure of quality.

Bandits Warm-up Cold Recommender Systems [35]

We address the cold start problem in recommendation systems assuming no contextual information is available neither about users, nor items. We consider the case in which we only have access to a set of ratings of items by users. Most of the existing works consider a batch setting, and use cross-validation to tune parameters. The classical method consists in minimizing the root mean square error over a training subset of the ratings which provides a factorization of the matrix of ratings, interpreted as a latent representation of items and users. Our contribution in this paper is 5-fold. First, we explicit the issues raised by this kind of batch setting for users or items with very few ratings. Then, we propose an online setting closer to the actual use of recommender systems; this setting is inspired by the bandit framework. The proposed methodology can be used to turn any recommender system dataset (such as Netflix, MovieLens,...) into a sequential dataset. Then, we explicit a strong and insightful link between contextual bandit algorithms and matrix factorization; this leads us to a new algorithm that tackles the exploration/exploitation dilemma associated to the cold start problem in a strikingly new perspective. Finally, experimental evidence confirm that our algorithm is effective in dealing with the cold start problem on publicly available datasets. Overall, the goal of this paper is to bridge the gap between recommender systems based on matrix factorizations and those based on contextual bandits.

6.2.4. Nonparametric statistics of time series

Uniform hypothesis testing for finite-valued stationary processes [6]

Given a discrete-valued sample $X_1, \cdots, X_n$ we wish to decide whether it was generated by a distribution belonging to a family H_0, or it was generated by a distribution belonging to a family H_1. In this work we assume that all distributions are stationary ergodic, and do not make any further assumptions (e.g. no independence or mixing rate assumptions). We would like to have a test whose probability of error (both Type I and Type II) is uniformly bounded. More precisely, we require that for each ϵ there exist a sample size n such that probability of error is upper-bounded by ϵ for samples longer than n. We find some necessary and some sufficient conditions on H_0 and H_1 under which a consistent test (with this notion of consistency) exists. These conditions are topological, with respect to the topology of distributional distance.

Asymptotically consistent estimation of the number of change points in highly dependent time series [17]

The problem of change point estimation is considered in a general framework where the data are generated by arbitrary unknown stationary ergodic process distributions. This means that the data may have long-range dependencies of an arbitrary form. In this context the consistent estimation of the number of change points is provably impossible. A formulation is proposed which overcomes this obstacle: it is possible to find the correct number of change points at the expense of introducing the additional constraint that the correct number of process distributions that generate the data is provided. This additional parameter has a natural interpretation in many real-world applications. It turns out that in this formulation change point estimation can be reduced to time series clustering. Based on this reduction, an algorithm is proposed that finds the number of change points and locates the changes. This algorithm is shown to be asymptotically consistent. The theoretical results are complemented with empirical evaluations.

6.3. Statistical Learning and Bayesian Analysis

6.3.1. Prediction of Sequences of Structured and Unstructured Data

Statistical performance analysis of a fast super-resolution technique using noisy translations [38]

It is well known that the registration process is a key step for super-resolution reconstruction. In this work, we propose to use a piezoelectric system that is easily adaptable on all microscopes and telescopes for controlling accurately their motion (down to nanometers) and therefore acquiring multiple images of the same scene at different controlled positions. Then a fast super-resolution algorithm can be used for efficient super-resolution reconstruction. In this case, the optimal use of r2 images for a resolution enhancement factor r is generally not enough to obtain satisfying results due to the random inaccuracy of the positioning system. Thus we propose to take several images around each reference position. We study the error produced by the super-resolution algorithm due to spatial uncertainty as a function of the number of images per position. We obtain a lower bound on the number of images that is necessary to ensure a given error upper bound with probability higher than some desired confidence level.

Quantitative control of the error bounds of a fast super-resolution technique for microscopy and astronomy [11]

While the registration step is often problematic for super-resolution, many microscopes and telescopes are now equipped with a piezoelectric mechanical system which permits to ac-curately control their motion (down to nanometers). There-fore one can use such devices to acquire multiple images of the same scene at various controlled positions. Then a fast super-resolution algorithm [1] can be used for efficient super-resolution. However the minimal use of r 2 images for a resolution enhancement factor r is generally not sufficient to obtain good results. We propose to take several images at po-sitions randomly distributed close to each reference position. We study the number of images necessary to control the error resulting from the super-resolution algorithm by [1] due to the uncertainty on positions. The main result is a lower bound on the number of images to respect a given error upper bound with probability higher than a desired confidence level.

6.3.2. Statistical analysis of superresolution

A diffusion strategy for distributed dictionary learning [12]

We consider the problem of a set of nodes which is required to collectively learn a common dictionary from noisy measurements. This distributed dictionary learning approach may be useful in several contexts including sensor networks. Dif-fusion cooperation schemes have been proposed to estimate a consensus solution to distributed linear regression. This work proposes a diffusion-based adaptive dictionary learning strategy. Each node receives measurements which may be shared or not with its neighbors. All nodes cooperate with their neighbors by sharing their local dictionary to estimate a common representa-tion. In a diffusion approach, the resulting algorithm corresponds to a distributed alternate optimization. Beyond dictionary learn-ing, this strategy could be adapted to many matrix factorization problems in various settings. We illustrate its efficiency on some numerical experiments, including the difficult problem of blind hyperspectral images unmixing.

6.4. Miscellaneous

6.4.1. Miscellaneous

Online Matrix Completion Through Nuclear Norm Regularisation [14]

It is the main goal of this paper to propose a novel method to perform matrix completion on-line. Motivated by a wide variety of applications, ranging from the design of recommender systems to sensor network localization through seismic data reconstruction, we consider the matrix completion problem when entries of the matrix of interest are observed gradually. Precisely, we place ourselves in the situation where the predictive rule should be refined incrementally, rather than recomputed from scratch each time the sample of observed entries increases. The extension of existing matrix completion methods to the sequential prediction context is indeed a major issue in the Big Data era, and yet little addressed in the literature. The algorithm promoted in this article builds upon the Soft Impute approach introduced in Mazumder et al. (2010). The major novelty essentially arises from the use of a randomised technique for both computing and updating the Singular Value Decomposition (SVD) involved in the algorithm. Though of disarming simplicity, the method proposed turns out to be very efficient, while requiring reduced computations. Several numerical experiments based on real datasets illustrating its performance are displayed, together with preliminary results giving it a theoretical basis.

Synthèse en espace et temps du rayonnement acoustique d'une paroi sous excitation turbulente par synthèse spectrale 2D+T et formulation vibro-acoustique directe [33]

Une méthode directe pour simuler les vibrations et le rayonnement acoustique d'une paroi soumise à un écoulement subsonique est proposée. Tout d'abord, en adoptant l'hypothèse d'un écoulement homogène et stationnaire, on montre qu'une méthode de synthèse spectrale en espace et temps (2D+t) est suffisante pour obtenir explicitement une réalisation d'un champ de pression pariétale excitatrice p(x,y,t) dont les propriétés inter-spectrales sont prescrites par un modèle empirique de Chase. Cette pression turbulente p(x,y,t) est obtenue explicitement et permet de résoudre le problème vibro-acoustique de la paroi dans une formulation directe. La méthode proposée fournit ainsi une solution complète du problème dans le domaine spatio-temporel : pression excitatrice, déplacement en flexion et pression acoustique rayonnée par la paroi. Une caractéristique de la méthode proposée est un cout de calcul qui s'avère similaire aux formulations inter-spectrales majoritairement utilisées dans la littérature. En particulier, la synthèse permet de prendre en compte l'intégralité des échelles spatio-temporelles du problème : échelles turbulentes, vibratoires et acoustiques. A titre d'exemple, la pression aux oreilles d'un auditeur suite à l'excitation turbulente de la paroi est synthétisée.

Bandits attack function optimization [27]

We consider function optimization as a sequential decision making problem under the budget constraint. Such constraint limits the number of objective function evaluations allowed during the optimization. We consider an algorithm inspired by a continuous version of a multi-armed bandit problem which attacks this optimization problem by solving the tradeoff between exploration (initial quasi-uniform search of the domain)

and exploitation (local optimization around the potentially global maxima). We introduce the so-called Simultaneous Optimistic Optimization (SOO), a deterministic algorithm that works by domain partitioning. The benefit of such an approach are the guarantees on the returned solution and the numerical eficiency of the algorithm. We present this machine learning rooted approach to optimization, and provide the empirical assessment of SOO on the CEC'2014 competition on single objective real-parameter numerical optimization testsuite.

Optimistic planning in Markov decision processes using a generative model [30]

We consider the problem of online planning in a Markov decision process with discounted rewards for any given initial state. We consider the PAC sample com-plexity problem of computing, with probability 1-δ, an -optimal action using the smallest possible number of calls to the generative model (which provides reward and next-state samples). We design an algorithm, called StOP (for Stochastic-Optimistic Planning), based on the "optimism in the face of uncertainty" princi-ple. StOP can be used in the general setting, requires only a generative model, and enjoys a complexity bound that only depends on the local structure of the MDP.

Near-Optimal Rates for Limited-Delay Universal Lossy Source Coding [3]

We consider the problem of limited-delay lossy coding of individual sequences. Here, the goal is to design (fixed-rate) compression schemes to minimize the normalized expected distortion redundancy relative to a reference class of coding schemes, measured as the difference between the average distortion of the algorithm and that of the best coding scheme in the reference class. In compressing a sequence of length T, the best schemes available in the literature achieve an $O(T^{-1/3})$ normalized distortion redundancy relative to finite reference classes of limited delay and limited memory, and the same redundancy is achievable, up to logarithmic factors, when the reference class is the set of scalar quantizers. It has also been shown that the distortion redundancy is at least of order $T^{-1/2}$ in the latter case, and the lower bound can easily be extended to sufficiently powerful (possibly finite) reference coding schemes. In this paper, we narrow the gap between the upper and lower bounds, and give a compression scheme whose normalized distortion redundancy is $O(ln(T)/T^{1/2})$ relative to any finite class of reference schemes, only a logarithmic factor larger than the lower bound. The method is based on the recently introduced shrinking dartboard prediction algorithm, a variant of exponentially weighted average prediction. The algorithm is also extended to the problem of joint source-channel coding over a (known) stochastic noisy channel and to the case when side information is also available to the decoder (the Wyner–Ziv setting). The same improvements are obtained for these settings as in the case of a noiseless channel. Our method is also applied to the problem of zero-delay scalar quantization, where $O(ln(T)/T^{1/2})$ normalized distortion redundancy is achieved relative to the (infinite) class of scalar quantizers of a given rate, almost achieving the known lower bound of order $1/T^{-1/2}$. The computationally efficient algorithms known for scalar quantization and the Wyner–Ziv setting carry over to our (improved) coding schemes presented in this paper.

Online Markov Decision Processes Under Bandit Feedback [4]

Software systems are composed of many interacting elements. A natural way to abstract over software systems is to model them as graphs. In this paper we consider software dependency graphs of object-oriented software and we study one topological property: the degree distribution. Based on the analysis of ten software systems written in Java, we show that there exists completely different systems that have the same degree distribution. Then, we propose a generative model of software dependency graphs which synthesizes graphs whose degree distribution is close to the empirical ones observed in real software systems. This model gives us novel insights on the potential fundamental rules of software evolution.

A Generative Model of Software Dependency Graphs to Better Understand Software Evolution [37]

Software systems are composed of many interacting elements. A natural way to abstract over software systems is to model them as graphs. In this paper we consider software dependency graphs of object-oriented software and we study one topological property: the degree distribution. Based on the analysis of ten software systems written in Java, we show that there exists completely different systems that have the same degree distribution. Then, we propose a generative model of software dependency graphs which synthesizes graphs whose degree

distribution is close to the empirical ones observed in real software systems. This model gives us novel insights on the potential fundamental rules of software evolution.

Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows [8]

We address the problem of rank elicitation as-suming that the underlying data generating pro-cess is characterized by a probability distribu-tion on the set of all rankings (total orders) of a given set of items. Instead of asking for complete rankings, however, our learner is only allowed to query pairwise preferences. Using information of that kind, the goal of the learner is to reliably predict properties of the distribution, such as the most probable top-item, the most probable rank-ing, or the distribution itself. More specifically, learning is done in an online manner, and the goal is to minimize sample complexity while guaran-teeing a certain level of confidence.

Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm [1]

We introduce a novel approach to preference-based reinforcement learn-ing, namely a preference-based variant of a direct policy search method based on evolutionary optimization. The core of our approach is a preference-based racing algorithm that selects the best among a given set of candidate policies with high probability. To this end, the algorithm operates on a suitable ordinal preference structure and only uses pairwise comparisons between sample rollouts of the policies. Embedding the racing algorithm in a rank-based evolutionary search procedure, we show that approxima-tions of the so-called Smith set of optimal policies can be produced with certain theoretical guarantees. Apart from a formal performance and complexity analysis, we present first experimental studies showing that our approach performs well in practice.

Biclique Coverings, Rectifier Networks and the Cost of ε-Removal [16]

We relate two complexity notions of bipartite graphs: the minimal weight biclique covering number Cov(G) and the minimal rec-tifier network size Rect(G) of a bipartite graph G. We show that there exist graphs with Cov(G) $\geq$ Rect(G) 3/2-ϱ . As a corollary, we estab-lish that there exist nondeterministic finite automata (NFAs) with ε-transitions, having n transitions total such that the smallest equivalent ε-free NFA has Ω(n 3/2-ϱ) transitions. We also formulate a version of previous bounds for the weighted set cover problem and discuss its con-nections to giving upper bounds for the possible blow-up.

Efficient Eigen-updating for Spectral Graph Clustering [2]

Partitioning a graph into groups of vertices such that those within each group are more densely connected than vertices assigned to different groups, known as graph clustering, is often used to gain insight into the organisation of large scale networks and for visualisation purposes. Whereas a large number of dedicated techniques have been recently proposed for static graphs, the design of on-line graph clustering methods tailored for evolving networks is a challenging problem, and much less documented in the literature. Motivated by the broad variety of applications concerned, ranging from the study of biological networks to the analysis of networks of scientific references through the exploration of communications networks such as the World Wide Web, it is the main purpose of this paper to introduce a novel, computationally efficient, approach to graph clustering in the evolutionary context. Namely, the method promoted in this article can be viewed as an incremental eigenvalue solution for the spectral clustering method described by Ng. et al. (2001). The incremental eigenvalue solution is a general technique for finding the approximate eigenvectors of a symmetric matrix given a change. As well as outlining the approach in detail, we present a theoretical bound on the quality of the approximate eigenvectors using perturbation theory. We then derive a novel spectral clustering algorithm called Incremental Approximate Spectral Clustering (IASC). The IASC algorithm is simple to implement and its efficacy is demonstrated on both synthetic and real datasets modelling the evolution of a HIV epidemic, a citation network and the purchase history graph of an e-commerce website.

From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning [36]

This work covers several aspects of the optimism in the face of uncertainty principle applied to large scale optimization problems under finite numerical budget. The initial motivation for the research reported here originated from the empirical success of the so-called Monte-Carlo Tree Search method popularized in computer-go and further extended to many other games as well as optimization and planning problems. Our objective is to contribute to the development of theoretical foundations of the field by characterizing the complexity of the underlying optimization problems and designing efficient algorithms with performance guarantees. The main idea presented here is that it is possible to decompose a complex decision making problem (such as an optimization problem in a large search space) into a sequence of elementary decisions, where each decision of the sequence is solved using a (stochastic) multi-armed bandit (simple mathematical model for decision making in stochastic environments). This so-called hierarchical bandit approach (where the reward observed by a bandit in the hierarchy is itself the return of another bandit at a deeper level) possesses the nice feature of starting the exploration by a quasi-uniform sampling of the space and then focusing progressively on the most promising area, at different scales, according to the evaluations observed so far, and eventually performing a local search around the global optima of the function. The performance of the method is assessed in terms of the optimality of the returned solution as a function of the number of function evaluations. Our main contribution to the field of function optimization is a class of hierarchical optimistic algorithms designed for general search spaces (such as metric spaces, trees, graphs, Euclidean spaces, ...) with different algorithmic instantiations depending on whether the evaluations are noisy or noiseless and whether some measure of the "smoothness" of the function is known or unknown. The performance of the algorithms depend on the local behavior of the function around its global optima expressed in terms of the quantity of near-optimal states measured with some metric. If this local smoothness of the function is known then one can design very efficient optimization algorithms (with convergence rate independent of the space dimension), and when it is not known, we can build adaptive techniques that can, in some cases, perform almost as well as when it is known.

<p style="text-align:center">SPIRALS Team</p>

6. New Results

6.1. Highlights of the Year

In 2014, we are proud to have organized the 17th ACM SIGSOFT International Conference on Component-Based Software Engineering and Software Architecture (CompArch) that has been held in Lille from 30 June to 3 July 2014.

CompArch is the main conference of the ACM SIGSOFT group on software architectures and software components. The conference is held alternatively in North America and in Europe. The 17th edition has been held this year in France for the first time. The conference brings together about 100 researchers from the academia and the industry.

6.2. Distributed Context Monitoring

In 2014, we obtained some new results in the area of distributed context monitoring solutions to support the development of self-optimising software systems. Context monitoring has emerged as a key capability in various domains to connect software systems to the underlying hardware platform or to the physical world (in the case of ubiquitous systems). In particular, we have investigated to the capability of inferring high-level contextual situations from a large volume of raw data collected from a single device or in the wild. Both hardware (*e.g.*, accelerometer) or software (*e.g.*, performance counters) sensors tend to continuously produce raw data that a context monitoring solution has to quickly filter, process, and convert it into information that can be used by an application or understood by a user.

As a result of the PhD thesis of Adel Noureddine [14], defended in March 2014, we have developed a middleware toolkit to support *in-depth context monitoring* in the domain of green computing. In particular, we introduce a software library, named POWERAPI, that can estimate the power consumption in real-time at various granularities of software: from system processes to code methods (see Section 5.3). This non-invasive solution provides accurate insights on energy hotspots of software and can be used to derive the energy profile of any software library, thus guiding the developers in optimising the energy consumption of their developments.

As a result of the PhD thesis of Nicolas Haderer [12], defended in November 2014, we have contributed to the development of a middleware platform to support *in-breadth context monitoring* in the area of mobile computing. In particular, we promote the distributed middleware solution APISENSE® as an efficient approach to deploy mobile crowd-sensing tasks across a large population of volunteer participants (see Section 5.1). In particular, APISENSE® includes a task orchestration algorithm that preserves the privacy and the battery of sensing devices, while maintaining specific sensing coverage objectives (including time and space dimensions). The server-side infrastructure of APISENSE® is generated from a dedicated software product line, while the implementation is based on the FRASCATI platform (see Section 5.2).

6.3. Design and Runtime Support for Cloud Computing

In 2014, we obtained some new results in the domain both of the design and the runtime support of distributed applications for multi-cloud systems. The purpose is to deal with applications that span across several different cloud systems. Several reasons justify such a goal. For example, in order to avoid the so-called vendor lock-in syndrome, cloud application stakeholders need to be able to migrate as easily as possible their assets from one cloud system to another one. Other examples include the possibility of introducing diversify and fault-tolerance by deploying applications on different cloud systems, or hot migrating applications where computing resources are less expensive.

For the design of multi-cloud systems, we proposed a solution based on software product lines (SPL) [90] and ontologies. In order to specify the variability of such environments, we extended SPL with attributes, cardinalities, and constraints. In order to enable the evolution of these environments, we provided an automated support for maintaining the consistency based on constraint programming. Finally, we proposed an ontology based approach to bridge the gap between the concepts and artefacts defined by different cloud systems. This global solution is the result of the PhD thesis of Clément Quinton [16] that was defended in October 2014, and has been partially supported by the FP7 PaaSage project (see Section 8.3).

For the runtime support of multi cloud systems, we proposed the SOCLOUD platform. This solution enables to deploy, execute and manage an application that spans on several different cloud systems. SOCLOUD tackles the challenges of portability, provisioning, elasticity, and high availability. SOCLOUD defines a component-based and service-oriented architecture that provides an unified view of a set of cloud systems. SOCLOUD is the result of the PhD thesis of Fawaz Paraiso [15] that was defended in June 2014. SOCLOUD is implemented on top of the FRASCATI platform (see Section 5.2).

6.4. Extraction and Analysis of Knowledge for Automatic Software Repair

Automated software repair aims at assisting developers in order to improve the quality of software systems, for example by recommending some repair actions to fix bugs. Matias Martinez has presented in his PhD thesis [13] that was defended in June 2014, new results in this domain. These results aim at reducing the search space when repairing a software system. The solution relies on two techniques. The first one consists in building change models learnt from repairs performed by other developers. These repairs are mined from existing software repositories of open source projects, and analysed based on their types and frequencies. The second proposed technique is based on the inherent redundancy of code patterns. The assumption is that the probability that the repair code for a particular kind of defect is already present in the software system under study is high. We then take advantage of this inherent redundancy to reduce the search space when looking for repair actions.