RESEARCH CENTER
Lille - Nord Europe

FIELD

Activity Report 2018

Section New Results

Edition: 2019-03-07

BONSAI Project-Team

6. New Results

6.1. Exploration of transcriptomes

In 2016 we produced a method called CG-Alcode able to compare transcripts repertoires of a given pair of orthologous genes. We applied our method to compare human and mouse transcriptomes. This year, in collaboration with C.Belleannée (DYLISS, Inria Rennes) we explored the comparison of multiple species. We inspected human, mouse and dog transcriptomes. We thus were able to predict a large number of putative transcripts in both human, mouse and dog based on known transcripts. Those results allow to investigate which functional sites are conserved and which genes have the same set of transcripts (known or putative).

6.2. Modeling of alternative transcripts with long reads

In the context of transcriptomic analyses based on third generation sequencing data (ONT), we started to explore the following problem : given a transcriptomic experiment, a gene of interest, select reads related to the given gene and find exon junctions. As we have done in the CG-Alcode project, we aim to model the gene as an alphabet of exonic blocks, each transcript being a word over this alphabet. This work takes place in the context of ANR ASTER for which we deal with mouse transcriptomic data in brain and liver. Built models will allow to query human genes to discover putative transcripts.

6.3. Read against read comparison for Nanopore data

In the team, we developed two years ago seeds with errors, which allow to find all common approximate patterns with a limited number of errors. The idea behind these seeds, called 01^*0 seeds, is to divide the sequence in blocks so that the distribution of errors is no longer random. This year, we have used these seeds in the context of long reads analysis. With this data, reads against reads comparison suffers from a high loss of sensitivity, because the single *read error-rate* is already high. Our application case is the detection of adapter sequences in ONT sequencing. We have shown that the use of these seeds instead of exact k-mers allowed a more accurate reconstruction of the sequences of the adapters. The method takes two steps: first the identification of k-mers potentially composing the adapter using a counting approach that takes into account errors in the read, and then the reconstruction of the complete sequence of the adapter with a greedy algorithm. Our results show that the seeds with errors allow to obtain accurate consensus sequences for more 80% of the samples, compared to 40% with the usual k-mer approach. This work was done within the ANR ASTER during the first year of the thesis of Quentin Bonenfant and was presented at the national workshop Seqbio 2018.

6.4. Annotation of the OC43 coronavirus genome

OC43 coronavirus is recognized as frequent cause of respiratory infection. We have conducted a bioinformatics study of 8 coronavirus genomes collected from patients at Lille hospital : gene annotation, phylogenetic analysis and amino acids substitutions. Several genotypes (B, E, F and G) were identified and two clusters of patients were defined from chronological data and phylogenetic trees based on the genomic sequences,. Analyses of amino acids substitutions of the S protein sequences identify substitutions specific of genotype F strains circulating among French people. This work is a collaboration with Anne Goffard (CHRU Lille and CIIL).

6.5. Small RNAs catalog in oilseed rape

Polyploidy – and notably allopolyploidy that involves interspecific hybridization – has played a major role in the evolution of plants, partly because this process is often associated with genomic structure and expression changes. Homeologous exchanges (HE) – i.e. between the constituent subgenomes– have been demonstrated to be frequent in allopolyploids and could be involved in the origin and maintenance of polyploids. While their influence on gene content has poorly been studied until recently, little is known about their impact on gene expression. Together with K. Alix (Inra Moulon), we have analyzed the impact of HEs that have been characterized in resynthesized oilseed rapes, on the repertoire of micro RNAs. Our main objective was to assess the relations that could exist between structural variation and modifications of gene expression through changes in miRNA regulation. The analysis was based on the small RNA-seq catalog obtained with the bioinformatic tool miRkwood, developed in BONSAI. We have built a microRNA database for the diploid subgenomes AA from *Brassica rapa* and CC for *Brassica oleracea* that correspond to the progenitors of the resynthesized *Brassica napus* allotetraploids (AACC). Integrating miRNA prediction and genomic location of HEs allowed us to infer relationships between microRNA restructuring and non-additivity of gene expression in polyploid hybrids.

6.6. Identifying systematic sequencing errors

Discovering over-represented approximate motifs in DNA sequences is an essential part of bioinformatics, which has been studied extensively. However, it remains a difficult challenge, especially with the huge quantity of data generated by high throughput sequencing technologies. We have developed an exact discriminative method for IUPAC motifs discovery in large sets of DNA sequences. The approach uses mutual information (MI) as an objective function to search for over-represented degenerate motifs in a lattice [7].

The algorithm was applied to the problem of *Sequence-Specific Errors*. Next Generation Sequencing, and further Single-Molecule Sequencing technologies are known to produce a highly variable error rate. A common method to overcome these sequencing errors is to increase the *coverage*. However, Sequence-Specific Errors are recurrent errors that depend on the upstream nucleotidic context, and can thus be confused with true genomic variations when the read coverage increases. Our algorithm was able to find motifs associated to sequencing errors and therefore to improve variant calling. This method has also tested on ChIP-seq datasets, and compared with five state-of-the art methods, where it was experimentally shown to perform as well as the best one, while be resistant to down-sampling.

This work was done during the thesis of Chadi Saad, and as a collaboration with Martin Figeac (Univ. Lille - Plateau de génomique fonctionnelle et structurale), Julie Leclerc and Marie-Pierre Buisine (CHRU de Lille - JPARC), and Hugues Richard (Sorbonne Université - Laboratory Computational and Quantitative Biology).

6.7. Indexing labelled sequences

We designed a compressed full-text index structure able to index a whole text with labels attached to every letter in the text [6]. This work will be applied to DNA sequences and more precisely V(D)J recombinations which are complex genomic rearrangements occurring in lymphocytes. The index will be used to index labelled V(D)J recombinations, which are labelled with their V, D and J gene. As the index we conceived is scalable, we will index V(D)J recombinations from thousands of samples and give access to this data through the Vidjil platform.

6.8. Tree representations

We found an intriguing duality between two well-known representations of trees [12]. This work concerns data structures and succinct tree representations. The Balanced Parenthesis representation of trees consists of encoding the structure of any tree using a series of opening and closing parentheses. The DFUDS representation is similar, but differs in how each node is encoded (also using parentheses). By relating both BP and DFUDS representations, we obtained improvements for a basic fundamental problem: the Minimum Length Interval Query problem. We also reported unnoted commonalities in recent solution to the Range Minimum Query problem.

6.9. Co-linear chaining on graphs

We reported the first algorithm that perform co-linear chaining between a sequence and a directed acyclic graph (DAG) [9]. This work concerns dynamic programming algorithms and sequencing alignment. The problem of co-linear chaining is a classical bioinformatics problem, which has immediate application to sequence alignment, as it is used as a filter remove spurious alignment seeds. Co-linear chaining is typically solved using a simple dynamic programming algorithm. Yet, representations of genomes using graphs instead of sequences have recently became an active research topic. As a results, the problem of aligning a sequence to a sequence graph merits consideration. This work provides the first step towards tackling practical sequence-to-graph alignment instances, by first considering the case when the graph is a DAG. We designed a $O(k|E|log|V|)$ algorithm to solve co-linear chaining on DAGs, which matches the optimal solution for the classical sequence variant, i.e. when the graph is a path.

6.10. Representations of de Bruijn graphs

We designed the first practical data structure for representing large de Bruijn graphs, which supports insertions and deletions of nodes [3]. This work concerns *de novo* assembly and several other k-mer-related bioinformatics problems. The representation of de Bruijn graphs is a transversal bioinformatics question that has enjoyed recent applications in genome, metagenome and transcriptome assembly and quantification. To this date, efficient data structures were essentially static. In this work we provided an implementation of a dynamic data structure that combines perfect hashing, Karp-Rabin hashing, and forests. Practical tests show that this structure is highly competitive with the state of the art.

6.11. Readability of overlap graphs

We report further progress on the study of a theoretical parameter of graph named *readability* [8]. This work concerns graph theory mainly. The readability parameter measures the minimal length of strings that would be needed in order to label a graph such that it is an overlap graph over a set of strings of that length. So far, recent works on readability have not elucidated many aspects related to this parameter: the complexity of computing it is open, and it is not even known whether the corresponding decision problem is in NP. The only upper bound known for this parameter is exponential. This work focuses on certain graph families: bipartite chain graphs, grids, induced subgraphs of grids, and provides a characterization of bipartite graphs of readability 2.

6.12. Nonribosomal peptides

Norine is a comprehensive public database for non-ribosomal peptides developed by the team for more than 10 years. The Norine database quality has been enhanced through a semi-automatic curation process of data. Particularly, more than 500 SMILES annotations have been added or updated. This allowed us to check and correct the monomeric graphs, i.e. a 2D representation of the monomeric composition of the NRPs, thanks to dedicated tools like Smiles2Monomers. This update was done in collaboration with members of the Proteome Informatics Group from SIB (Swiss Institute of Bioinformatics). New annotations on monoisotopic mass and molecular formulas have also been added. The Norine interface was improved and new features are available, such as the possibility to access the complete change history of each entry. To encourage new submissions of NRPs, authors of new NRPs are now visible as contributors on Norine home page. Finally, we published this year, in the field of biocontrol (a contraction of “biological control”), a paper on bioinformatic tools for the discovery of new lipopeptides [5], essentially based on the Norine platform.

BONUS Team

7. New Results

7.1. Decomposition-based optimization

- **A set-oriented decomposition algorithm for multi-objective optimization.**
  Participants: B. Derbel and A. Liefooghe, external collaborators: H. Aguirre and K. Tanaka, Shinshu University (JAPAN); S. Verel, Univ. Littoral (FRANCE); Q. Zhang, City University (HONG KONG)

  The working principles of the well-established multi-objective evolutionary algorithm MOEA/D relies on the iterative and cooperative improvement of a number of single-objective sub-problems obtained by decomposition. Besides the definition of sub-problems, selection and replacement are, like in any evolutionary algorithm, the two core elements of MOEA/D. We argue that these two components are however loosely coupled with the maintained population. Thereby, in [24], we propose to re-design the working principles of MOEA/D by adopting a set-oriented perspective, where a many-to-one mapping between sub-problems and solutions is considered. Selection is then performed by defining a neighborhood relation among solutions in the population set, depending on the corresponding sub-problem mapping. Replacement is performed following an elitist mechanism allowing the population to have a variable, but bounded, cardinality during the search process. By conducting a comprehensive empirical analysis on a range of combinatorial multi- and many-objective nk-landscapes, we show that the proposed approach leads to significant improvements, especially when dealing with an increasing number of objectives. Our findings indicate that a set-oriented design can constitute a sound alternative for strengthening the practice of multi- and many-objective evolutionary optimization based on decomposition.
- **Parallel Pareto local search for multi-objective optimization.**
  Participants: B. Derbel and A. Liefooghe, external collaborators: J. Shi and J. Sun, Xi'an Jiaotong University (CHINA); Q. Zhang, City University (HONG KONG)

  Pareto Local Search (PLS) is a simple, yet effective optimization approach dedicated to multi-objective combinatorial optimization. It can however suffer from a high computational cost, especially when the size of the Pareto optimal set is relatively large. Recently, incorporating decomposition in PLS had revealed a high potential, not only in providing high-quality approximation sets, but also in speeding-up the search process. In [30], using the bi-objective Unconstrained Binary Quadratic Programming (bUBQP) problem as an illustrative benchmark, we demonstrate some shortcomings in the resulting decomposition-guided Parallel Pareto Local Search (PPLS), and we propose to revisit the PPLS design accordingly. For instances with a priori unknown Pareto front shape, we show that a simple pre-processing technique to estimate the scale of the Pareto front can help PPLS to better balance the workload. Furthermore, we propose a simple technique to deal with the critically-important scalability issue raised by PPLS when deployed over a large number of computing nodes. Our investigations show that the revisited version of PPLS provides a consistent performance, suggesting that decomposition-guided PPLS can be further generalized in order to improve both parallel efficiency and approximation quality.
- **Archivers for the representation of the set of approximate solutions for MOPs.**
  Participants: E-G. Talbi, external collaborators: O. Schutze, C. Hernandez (Computer Science Department, Cinvestav, MEXICO), Q. Sun, Y. Naranjani (School of Engineering University of California, USA), R. Xiong (Department of Mechanics, University Tianjin, CHINA)

  In this work we have addressed the problem of computing suitable representations of the set of approximate solutions of a given multi-objective optimization problem via stochastic search algorithms. For this, we have proposed different archiving strategies for the selection of the candidate

solutions maintained by the generation process of the stochastic search process, and investigate them further on analytically and empirically. For all archivers we have provided upper bounds on the approximation quality as well as on the cardinality of the limit solution set. A comparative study on some test problems in order to visualize the effect of all novel archiving strategies has also been carried out [18].

7.2. ML-assisted optimization

FIve major contributions related to ML-assisted optimization have been achieved and summarized in the following. As pointed out previously in our research program, one of the major issues in surrogate-assisted optimization is how to integrate efficiently and effectively the surrogates in the optimization process. This issue is addressed in first three contributions. Another major aspect addressed in the fourth contribution is the investigation of surrogates within the context of combinatorial optimization. The focus of the fifth contribution is put on the landscape analysis applied within the context of multi-objective optimization.

- **Efficient Global Optimization Using Deep Gaussian Processes.**
  Participants: A. Hebbal, E-G. Talbi and N. Melab, external collaborators: L. Brevault and M. Balesdent from ONERA (Palaiseau, Paris)

  Efficient Global Optimization (EGO) is widely used for the optimization of computationally expensive black-box functions. EGO is based on a surrogate modeling technique using Gaussian Processes (kriging). However, due to the use of a stationary covariance, kriging is not well suited for approximating non stationary functions. Non stationarity is generally due to the abrupt change of a physical property that often occurs in the design of launch vehicles, subject of our collaboration with ONERA. This leads to a variation of the objective function with a completely different smoothness along the input space. In the spirit of deep learning using neural networks, we have investigated in [25] the integration of Deep Gaussian processes (DGP) in EGO framework to deal with non stationarity. Numerical experimentations are performed on analytical problems to highlight the different aspects of DGP and EGO. The experimental results show that the coupling EGO-DGP outperforms EGO-GP with a significant margin. Furthermore, the study has also highlighted some challenging issues to be investigated including: the integration of DGP in multi-objective EGO, the configuration of the network and revisiting the training model. Ultra-scale optimization at different levels is particularly important given the large number of hyperparameters of the training model.

- **Efficient global optimization of constrained mixed variable problems.**
  Participants: E-G. Talbi, external collaborators: Julien Pelamatti, Loïc Brevault, Mathieu Balesdent (ONERA) Yannick Guerin (CNES)

  Due to the increasing demand for high performance and cost reduction within the framework of complex system design, numerical optimization of computationally costly problems is an increasingly popular topic in most engineering fields [33]. In this work, several variants of the Efficient Global Optimization algorithm for costly constrained problems depending simultaneously on continuous decision variables as well as on quantitative and/or qualitative discrete design parameters are proposed. The adaptation that is considered is based on a redefinition of the Gaussian Process kernel as a product between the standard continuous kernel and a second kernel representing the covariance between the discrete variable values. Several parameterizations of this discrete kernel, with their respective strengths and weaknesses, have been investigated. The novel algorithms are tested on a number of analytical test-cases and an aerospace related design problem, and it is shown that they require fewer function evaluations in order to converge towards the neighborhoods of the problem optima when compared to more commonly used optimization algorithms [38].

- **Adaptive Evolution Control using Confident Regions for Surrogate-assisted Optimization.**
  Participants: G. Briffoteaux and N. Melab, external collaborators: M. Mezmaz and D. Tuyttens from Université de Mons (BELGIUM)

The challenge of the efficient/effective integration of surrogates in the optimization process is to find the best trade-off between the quality (in terms of quality/precision) of the generated solutions and the efficiency (in terms of execution time) of the resolution. In [22], we have investigated the evolution control that alternates between the real function (simulator) and the surrogate within the multi-objective optimization process. We propose an adaptive evolution control mechanism based on the distance-based concept of confident regions (hyperspheres). The approach has been integrated into an ANN-assisted NSGA-2 and experimented using the ZDT4 multi-modal benchmark function. The reported results show that the proposed approach outperforms two other existing ones.

- **A surrogate model for combinatorial optimization.**
  Participants: B. Derbel and A. Liefooghe, external collaborators: H. Aguirre and K. Tanaka, Shinshu University (JAPAN), S. Verel, Univ. Littoral (FRANCE)

  Extensive efforts so far have been devoted to the design of effective surrogate models for expensive black-box continuous optimization problems. There are, however, relatively few investigations on the development of methodologies for combinatorial domains. In [31], we rely on the mathematical foundations of discrete Walsh functions in order to derive a surrogate model for pseudo-boolean optimization functions. Specifically, we model such functions by means of Walsh expansion. By conducting a comprehensive set of experiments on nk-landscapes, we provide empirical evidence on the accuracy of the proposed model. In particular, we show that a Walsh-based surrogate model can outperform the recently-proposed discrete model based on Kriging.

- **Landscape analysis for multi-objective optimization.**
  Participants: B. Derbel and A. Liefooghe, external collaborators: H. Aguirre and K. Tanaka, Shinshu University (JAPAN); M. López-Ibánez, Univ. Manchester (UK); L. Paquete, Univ. Coimbra, Portugal; S. Verel, Univ. Littoral (FRANCE)

  Pareto local optimal solutions (PLOS) are believed to highly influence the dynamics and the performance of multi-objective optimization algorithms, especially those based on local search and Pareto dominance. In [28], we introduce a PLOS network (PLOS-net) model as a step toward the fundamental understanding of multi-objective landscapes and search algorithms. Using a comprehensive set of instances, PLOS-nets are constructed by full enumeration, and selected network features are further extracted and analyzed with respect to instance characteristics. A correlation and regression analysis is then conducted to capture the importance of the PLOS-net features on the runtime and effectiveness of two prototypical Pareto-based heuristics. In particular, we are able to provide empirical evidence for the relevance of the PLOS-net model to explain algorithm performance.

  Additionally, we know that local search algorithms naturally stop at a local optimal set (LO-set) under given definitions of neighborhood and preference relation among subsets of solutions, such as set-based dominance relation, hypervolume or epsilon indicator. It is an open question how LO-sets under different set preference relations relate to each other. In [29], we report an in-depth experimental analysis on multi-objective nk-landscapes. Our results reveal that, whatever the preference relation, the number of LO-sets typically increases with the problem non-linearity, and decreases with the number of objectives. We observe that strict LO-sets of bounded cardinality under set-dominance are LO-sets under both epsilon and hypervolume, and that LO-sets under hypervolume are LO-sets under set-dominance, whereas LO-sets under epsilon are not. Nonetheless, LO-sets under set-dominance are more similar to LO-sets under epsilon than under hypervolume. These findings have important implications for multi-objective local search. For instance, a dominance-based approach with bounded archive gets more easily trapped and might experience difficulty to identify an LO-set under epsilon or hypervolume. On the contrary, a hypervolume-based approach is expected to perform more steps before converging to better approximations.

7.3. Large scale GPU-centric optimization

Participants: J. Gmys, T. C. Pessoa and N. Melab, external collaborators: M. Mezmaz, D. Tuyttens from University of Mons (BELGIUM) and F.H. De Carvalho Junior from Universidade Federal Do Ceará (BRAZIL)

Nowadays, accelerator-centric architectures offer orders-of-magnitude performance and energy improvements. The interest of those parallel resources has been recently accentuated by the advent of deep learning making them definitely key-building blocks of modern supercomputers. During the year 2018, in collaboration with A. Zomaya (The Univ. of Sydney) and I. Chakroun (IMEC, Leuven) N. Melab has (guest-)edited a special issue on this hot topic (editorial in [16]). In addition, we have put the focus on the investigation of these specific devices within the context of parallel optimization. In the following, two major contributions are reported: (1) Many-core Branch-and-Bound for GPU accelerators and MIC coprocessors; (2) Cuda Dynamic Parallelism (CDP) for backtracking.

- **Many-core Branch-and-Bound for GPU accelerators and MIC coprocessors.** Solving large optimization problems results in the generation of a very large pool of subproblems and the time-intensive evaluation of their associated lower bounds. Generating and evaluating those subproblems on coprocessors raises several issues including processor-coprocessor data transfer optimization, vectorization, thread divergence, etc. In [15], [32], we have investigated the offload-based parallel design and implementation of B&B algorithms for coprocessors addressing these issues. Two major many-core architectures are considered and compared: Nvidia GPU and Intel MIC. The proposed approaches have been experimented using the Flow-Shop scheduling problem and two hardware configurations equivalent in terms of energy consumption: Nvidia Tesla K40 and Intel Xeon Phi 5110P. The reported results show that the GPU-accelerated approach outperforms the MIC offload-based one even in its vectorized version. Moreover, vectorization improves the efficiency of the MIC offload-based approach with a factor of two.
- **Dynamic Configuration of CUDA Runtime Variables for CDP-based Divide-and-Conquer Algorithms.** CUDA Dynamic Parallelism (CDP) is an extension of the GPGPU programming model proposed to better address irregular applications and recursive patterns of computation. However, processing memory-demanding problems by using CDP is not straightforward, because of its particular memory organization. We have proposed in [23] (extension of [13]) an algorithm to deal with such an issue which dynamically calculates and configures the CDP runtime variables and the GPU heap on the basis of an analysis of the partial backtracking tree. We have implemented the algorithm for solving permutation problems and experimented on two test-cases: N-Queens and the Asymmetric Travelling Salesman Problem. The proposed algorithm allows different CDP-based backtracking from the literature to solve memory-demanding problems, adaptively with respect to the number of recursive kernel generations and the presence of dynamic allocations on GPU.

DEFROST Project-Team

7. New Results

7.1. Dynamic control of soft robots

The objective is to design a closed-loop strategy to control the dynamics of soft robots. We model the soft robot using the Finite Element Method, which leads to work with large-scale systems that are difficult to control. No unified framework exist to control these robots, especially when considering their dynamics. The main contribution of our work is a reduced order model-based control law, that consists in two main features: a reduced state feedback tunes the performance while a Lyapunov function guarantees the stability of the large-scale closed-loop systems. The method is generic and usable for any soft robot, as long as a FEM model is obtained. Simulation and real robots experiments show that we can control and reduce the settling time of the soft robot and make it converge faster without oscillations to a desired position. It can make the robot converge faster and with reduced oscillations to a desired equilibrium state in the robot's work-space. These results have been presented at the European Control Conference [24] and accepted for publication in Robotics and Automation Letters [8].

7.2. Vision-based force sensing for soft robots

This paper proposes a new framework of external force sensing for soft robots based on the fusion of vision-based measurements and Finite Element Model (FEM) techniques. A precise mechanical model of the robot is built using real-time FEM to describe the relationship between the external forces acting on the robot and the displacement of predefined feature points. The position of these feature points on the real robot is measured using a vision system and is compared with the equivalent feature points in the finite element model. Using the compared displacement, the intensities of the external forces are computed by solving an inverse problem. Based on the developed FEM equations, we show that not only the intensities but also the locations of the external forces can be estimated. A strategy is proposed to find the correct locations of external forces among several possible ones. The method is verified and validated using both simulation and experiments on a soft sheet and a parallel soft robot (both of them have non-trivial shapes). The good results obtained from the experimental study demonstrate the capability of our approach.

7.3. Fast, generic and reliable control and simulation of soft robots using model order reduction

Obtaining an accurate mechanical model of a soft deformable robot compatible with the computation time imposed by robotic applications is often considered as an unattainable goal. This paper should invert this idea. The proposed methodology offers the possibility to dramatically reduce the size and the online computation time of a Finite Element Model (FEM) of a soft robot. After a set of expensive offline simulations based on the whole model, we apply snapshot-proper orthogonal decomposition to sharply reduce the number of state variables of the soft robot model. To keep the computational efficiency, hyper-reduction is used to perform the integration on a reduced domain. The method allows to tune the error during the two main steps of complexity reduction. The method handles external loads (contact, friction, gravity...) with precision as long as they are tested during the offline simulations. The method is validated on two very different examples of FE models of soft robots and on one real soft robot. It enables acceleration factors of more than 100, while saving accuracy, in particular compared to coarsely meshed FE models and provides a generic way to control soft robots.

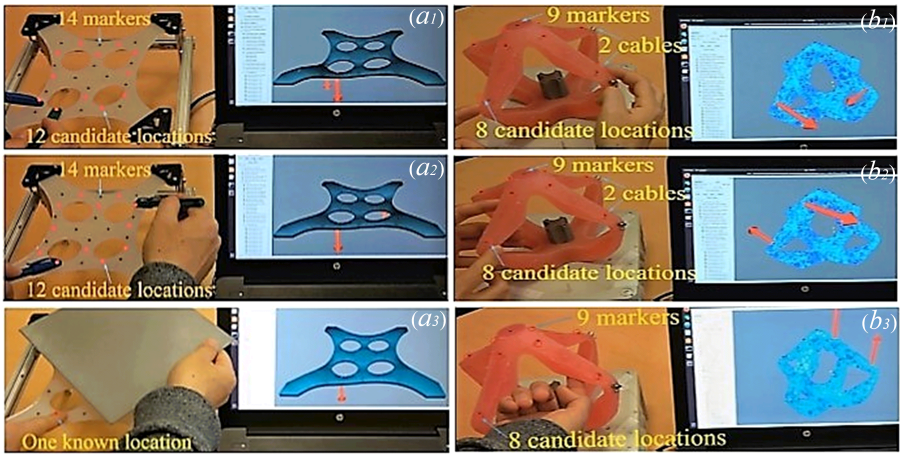

Figure 5. External force sensing for soft objects

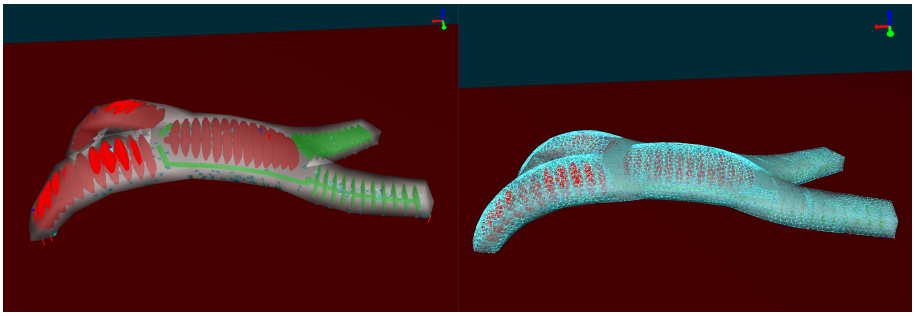

Figure 6. Pneumatic Soft Robot fine simulation versus its surrogate reduced representation manageable in real-time.

7.4. FEM-based kinematics and closed-loop control of soft, continuum manipulators

This paper presents a modeling methodology and experimental validation for soft manipulators to obtain forward and inverse kinematic models under quasistatic conditions. It offers a way to obtain the kinematic characteristics of this type of soft robots that is suitable for offline path planning and position control. The modeling methodology presented relies on continuum mechanics which does not provide analytic solutions in the general case. Our approach proposes a real-time numerical integration strategy based on Finite Element Method (FEM) with a numerical optimization based on Lagrangian Multipliers to obtain forward and inverse models. To reduce the dimension of the problem, at each step, a projection of the model to the constraint space (gathering actuators, sensors and end-effector) is performed to obtain the smallest number possible of mathematical equations to be solved. This methodology is applied to obtain the kinematics of two different manipulators with complex structural geometry. An experimental comparison is also performed in one of the robots, between two other geometric approaches and the approach that is showcased in this paper. A closed-loop controller based on a state estimator is proposed. The controller is experimentally validated and its robustness is evaluated using Lyapunov stability method.

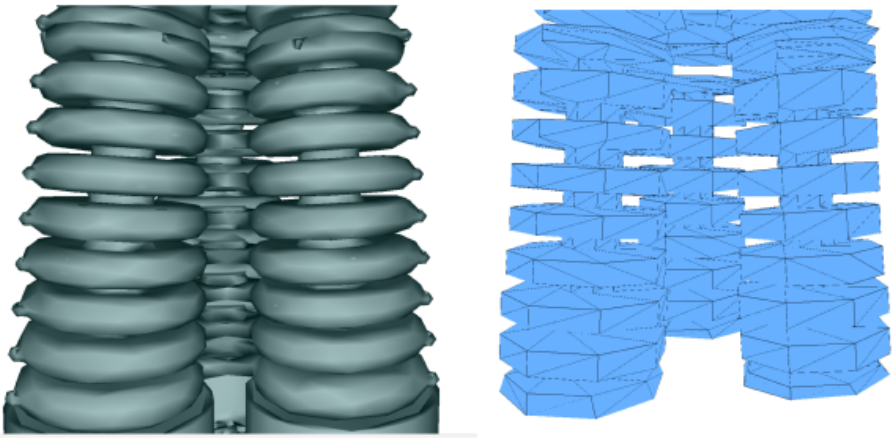

Figure 7. Visual model of the manipulator and the underlying finite element model.

7.5. FEM-based Deformation Control for Dexterous Manipulation of 3D Soft Objects

In this project, that was organize through a collaboration with Fanny Ficuciello from University of Naples and Antoine Petit from Mimesis team in Strasbourg we developed a method for dexterous manipulation of 3D soft objects for real-time deformation control, relying on Finite Element modelling. The goal is to generate proper forces on the fingertips of an anthropomorphic device during in-hand manipulation to produce desired displacements of selected control points on the object. The desired motions of the fingers are computed in real-time as an inverse solution of a Finite Element Method (FEM), the forces applied by the fingertips at the contact points being modelled by Lagrange multipliers. The elasticity parameters of the model are preliminarily estimated using a vision system and a force sensor. Experimental results were shown with an underactuated anthropomorphic hand that performs a manipulation task on a soft cylindrical object.

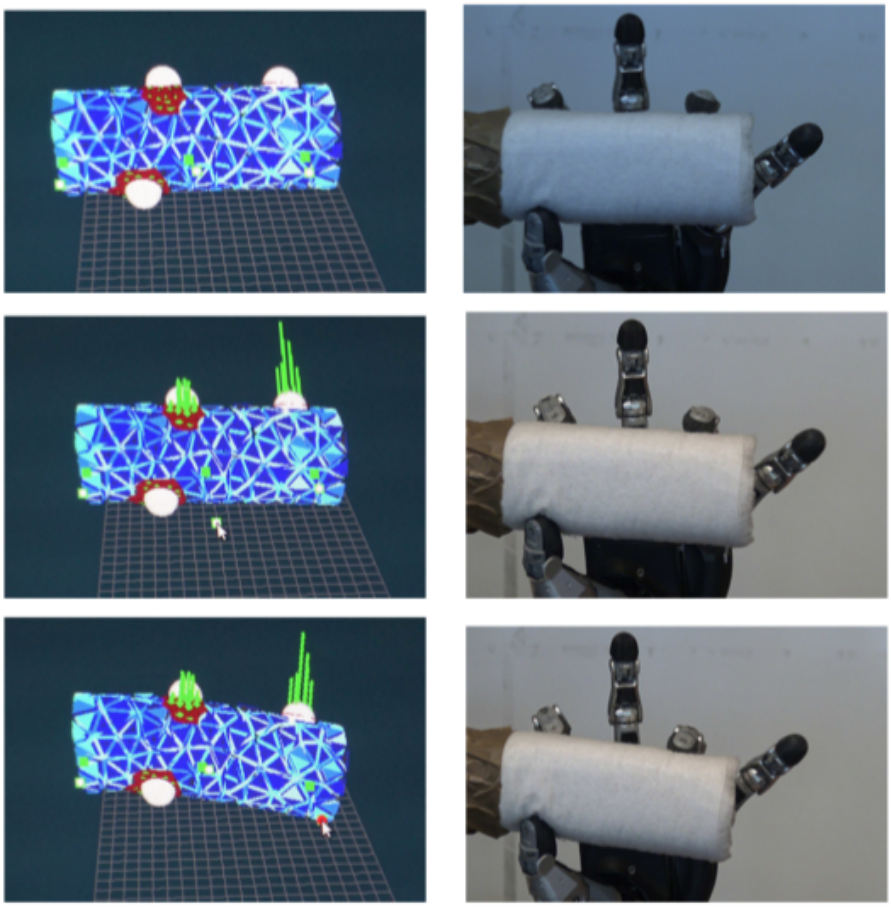

Figure 8. The manipulation of the 3D Soft Object inside the hand is driven by the inverse FEM simulation computed in real-time

FUN Project-Team

7. New Results

7.1. Performance Evaluation, Security, Safety and Verification

Participants: Antoine Gallais, Nathalie Mitton, Allan Blanchard.

7.1.1. Performance Evaluation and validation methodology

Envisioned communication densities in Internet of Things applications are increasing continuously. Because these wireless devices are often battery powered, we need specific energy efficient (low-power) solutions. Moreover, these smart objects use low-cost hardware with possibly weak links, leading to a lossy network. Once deployed, these low-power lossy networks (LLNs) are intended to collect the expected measurements, handle transient faults, topology changes, etc. Consequently, validation and verification during the protocol development are a matter of prime importance. A large range of theoretical or practical tools are available for performance evaluation. A theoretical analysis may demonstrate that the performance guarantees are respected, while simulations or experiments aim on estimating the behavior of a set of protocols within real-world scenarios. In [16], we review the various parameters that should be taken into account during such a performance evaluation. Our primary purpose is to provide a tutorial that specifies guidelines for conducting performance evaluation campaigns of network protocols in LLNs. We detail the general approach adopted in order to evaluate the performance of layer 2 and 3 protocols in LLNs. Furthermore, we also specify the methodology that should be adopted during the performance evaluation, while reviewing the numerous models and tools that are available to the research community.

7.1.2. Correlated failures

Current practices of fault-tolerant network design ignore the fact that most network infrastructure faults are localized or spatially correlated (i.e., confined to geo-graphic regions). Network operators require new tools to mitigate the impact of such region-based faults on their infrastructures. Utilizing the support from the U.S. Department of Defense, and by consolidating a wide range of theories and solutions developed in the last few years, [14] designs RAPTOR, an advanced Network Planning and Management Tool that facilitates the design and provisioning of robust and resilient networks.The tool provides multi-faceted network design, evaluation, and simulation capabilities for network planners. Future extensions of the tool currently being worked upon not only expand the tool's capabilities, but also extend these capabilities to heterogeneous interdependent networks such as communication, power, water, and satellite networks.

7.1.3. Contiki verification

Internet of Things (IoT) applications are becoming increasingly critical and require formal verification. Our recent work presented formal verification of the linked list module of Contiki, an OS for IoT. It relies on a parallel view of a linked list via a companion ghost array and uses an inductive predicate to link both views. In this work, a few interactively proved lemmas allow for the automatic verification of the list functions specifications, expressed in the acsl specification language and proved with the Frama-C/Wp tool. In a broader verification context, especially as long as the whole system is not yet formally verified, it would be very useful to use runtime verification , in particular, to test client modules that use the list module. It is not possible with the current specifications, which include an inductive predicate and axiomatically defined functions. In [27], an early-idea paper we show how to define a provably equivalent non-inductive predicate and a provably equivalent non-axiomatic function that belong to the executable subset e-acsl of acsl and can be transformed into executable C code. Finally, we propose an extension of Frama-C to handle both axiomatic specifications for deductive verification and executable specifications for runtime verification.

In [23], [47], we target Contiki, a widely used open-source OS for IoT, and present a verification case study of one of its most critical modules: that of linked lists. Its API and list representation differ from the classical linked list implementations, and are particularly challenging for deductive verification. The proposed verification technique relies on a parallel view of a list through a companion ghost array. This approach makes it possible to perform most proofs automatically using the Frama-C/WP tool, only a small number of auxiliary lemmas being proved interactively in the Coq proof assistant. We present an elegant segment-based reasoning over the companion array developed for the proof. Finally, we validate the proposed specification by proving a few functions manipulating lists.

With the wide expansion of multiprocessor architectures, the analysis and reasoning for programs under weak memory models has become an important concern. [13] presents MMFilter, an original constraint solver for generating program behaviors respecting a particular memory model. It is implemented in Prolog using CHR (Constraint Handling Rules). The CHR formalism provides a convenient generic solution for specifying memory models. It benefits from the existing optimized implementations of CHR and can be easily extended to new models. We present MMFilter design, illustrate the encoding of memory model constraints in CHR and discuss the benefits and limitations of the proposed technique.

7.2. Alternative communication paradigms

Participants: Antonio Costanzo, Valeria Loscri.

Nowadays, the always growing of connected objects and the strong demand to downsizing the devices in order to make the Internet of Things (IoT) paradigm more pervasive and ubiquitous, has motivated academic and industry people to investigate from one side mechanisms able to adapt quickly to the rapid external changes and to the quality of Services (QoS) parameters defined by the users and imposed by the adoption of new services and from another side, the investigation of portion of spectrum that have not been considered till this moment such as Terahertz band.

Nowadays, the always growing of connected objects and the strong demand to downsizing the devices in order to make the Internet of Things (IoT) paradigm more pervasive and ubiquitous, has motivated academic and industry people to investigate from one side mechanisms able to adapt quickly to the rapid external changes and to the quality of Services (QoS) parameters defined by the users and imposed by the adoption of new services and from another side, the investigation of portion of spectrum that have not been considered till this moment such as Terahertz band. In order to be able to realize a paradigm shift towards the Internet of Everything concept, a downsizing of devices is imposed allowing new applications as *in-vivo* diagnosis and monitoring. In order to be effective at this level it is imperative to analyze the new context, by highlighting the unique features to concretely realize the IoE paradigm. In this context, we have studied quantum particles called phonons, quasi-particles derived from vibrations of atoms in solids. Phonons have been envisaged as enabler of information transfer and their special characteristics have been exploited in [17]. Phonons have been also considered for a quantum channel in [26]. Another interesting approach for enabling the nano communication paradigm is represented by molecular communication. In particular, a main issue that is important to face is the coexistence between an artificial molecular communication and a biological system as explained in [40]. Alternative communication paradigms have attracted a lot of attention in the last a few years, not only by academic researchers but also by industry. Research on optical communication and in particular the possible exploitation of Visible Light communication with a twofold objective, to illuminate and to communicate has been object of an increasing interest. In this directions, we have proposed context-aware VLC systems in[39], [38] and [24]. The context is different in respect to the “traditional” wireless communication, since the external environment can change fastly and abruptly. Based on this primary observation, our main objective is to make theVLC system aware of the external noise and try to make it as robust as possible in respect of it.

7.3. Self-Organization

Participants: Antoine Gallais, Nathalie Mitton, Valeria Loscri, Farouk Mezghani, Anjalalaina Jean Cristanel Razafimandimby.

7.3.1. Stable parent selection

The Industrial Internet of Things consists in the use of low power lossy networks to enable next industrial applications. To work properly, the network has to provide strict guarantees concerning the delay and the reliability. IEEE 802.15.4-TSCH proposes time synchronized and slow channel hopping medium access control to cope with these requirements. It relies on a strict schedule of the transmissions, spread over orthogonal radio channels, to set up a resilient wireless infrastructure. A routing protocol (e.g. RPL) has then to construct energy-efficient routes on top of this link-layer topology (as investigated in the 6TiSCH IETF working group). Most of existing solutions rely on tree-based topologies, where each node has to select one or multiple parents to forward its traffic to the destination. Unfortunately, the links to the routing parents exhibit time-varying characteristics, due to e.g. obstacles, and external interference, thus leading to oscillations and increased required control of the routing topology. Moreover, the network has to provision enough resources (i.e., time, channel) to cope with those variations, while still being reactive to node/link failures. We investigated the stability of 6TiSCH networks, and especially the impact on routing parent selection. We identified moments of instability due to oscillations in the radio conditions caused by external interference and obstacles, in two indoor testbeds with different channel conditions. We identified the causes of instabilities, and proposed solutions for each of the layers in the 6TiSCH stack. First, at the MAC layer, we demonstrated that a rearrangement of shared cells in the slotframe reduces the probability of collisions for control packets, paving the way to a faster negotiation during topology reconfigurations. Next, we eased the schedule consistency management between two nodes (renegociated from scratch in the current standard, upon detection of a schedule inconsistency). Finally, at the routing layer, we exploited the existing correlation between the broadcast packet reception rate and the unicast link quality to create a two-step parent selection that favors stable parents. We finally obtained a network that converged faster and that reacted accurately during moments of instabilities. Results are available in [46], [42].

7.3.2. Bayesian communications

The amount of data that are generated in IoT devices is huge and the most of time data are highly correlated, by making useless the forwarding of all the raw data generated. Bearing that in mind, we have designated and implemented an effective mechanism to reduce the amount of data sent in the network in [45]. Results are encouraging since there is a size effect of less interfering in the communication system with an important impact on battery consumption for wireless devices that are energy constrained.

7.3.3. Multi-technology self-organization

Opportunistic communications present a promising solution for disaster network recovery in emergency situations such as hurricanes, earthquakes, and floods, where infrastructure might be destroyed. Some recent works in the literature have proposed opportunistic-based disaster recovery solutions, but they have omitted the consideration of mobile devices that come with different network technologies and various initial energy levels. [19], [30] present COPE, an energy-aware Cooperative OPportunistic alErt diffusion scheme for trapped survivors to use during disaster scenarios to report their position and ease their rescue operation. It aims to maintain mobile devices functional for as long as possible for maximum network coverage until reaching proximate rescuers. COPE deals with mobile devices that come with an assortment of networks and aims to perform systematic network interface selection. Furthermore, it considers mobile devices with various energy levels and allows low-energy nodes to hold their charge for longer time with the support of high-energy nodes. A proof-of-concept implementation has been performed to study the doability and efficiency of COPE, and to highlight the lessons learned. Following-up with these results, we performed several experimentations and could benchmark smartphone performances with regards to their multi-communications interfaces. Testing experiments have been carried out to measure the performance of smartphones in terms of energy consumption, clock synchronization and transmission range. We believe that such experimental results can support technological choices for rescue operations but also for many other applications relying on smartphone performances. Results are available in [30].

7.3.4. Heterogeneous Self-organizing (smart) Things

In the panorama of the Internet of Things, one main important issue is the management of heterogeneous objects, that need to communicate in order to exchange information and to interact in order to be able to synergically accomplish complex tasks and for providing services to final users. In this context, the thesis [10] has tried to face the main challenges related to complex heterogeneous systems, where objects are able to self-organize to each other and are equipped with some kind of intelligence in order to dynamically react to the environment changes. Several tools have been exploited ranging from artificial neural networks to genetic algorithms and different solutions have been proposed to make these systems dynamic and responding to the $*$self properties.

7.4. Smart Grids

Participants: Nathalie Mitton, Jad Nassar.

The Smart Grid (SG) aims to transform the current electric grid into a "smarter" network where the integration of renewable energy resources, energy efficiency and fault tolerance are the main benefits. This is done by interconnecting every energy source, storage point or central control point with connected devices, where heterogeneous SG applications and signaling messages will have different requirements in terms of reliability, latency and priority. Hence, data routing and prioritization are the main challenges in such networks.

So far, RPL (Routing Protocol for Low-Power and Lossy networks) protocol is widely used on Smart Grids for distributing commands over the grid. RPL assures traffic differentiation at the network layer in wireless sensor networks through the logical subdivision of the network in multiple instances, each one relying on a specific Objective Function. However, RPL is not optimized for Smart Grids, as its main objective functions and their associated metric does not allow Quality of Service differentiation.

In order to overcome this, we propose $OFQS$ an objective function [20] with a multi-objective metric that considers the delay and the remaining energy in the battery nodes alongside with the dynamic quality of the communication links. Our function automatically adapts to the number of instances (traffic classes) providing a Quality of Service differentiation based on the different Smart Grid applications requirements. We tested our approach on a real sensor testbed. The experimental results show that our proposal provides a lower packet delivery latency and a higher packet delivery ratio while extending the lifetime of the network compared to solutions in the literature.

The management of communication is an issue in WSN-based Smart Grid: billions of messages with different sizes and priorities are sent across the network. Data aggregation is a potential solution to reduce loads on the communication links, thus achieving a better utilization of the wireless channel and reducing energy consumption. On the other hand, SG applications require different Quality of Service (QoS) priorities. Delays caused by data aggregation must then be controlled in order to achieve a proper communication. In [33], [34], we propose a work in progress, that consists of a QoS efficient data aggregation algorithm with two aggregation functions for the different traffics in a SG network. We expect to reduce the energy consumption while respecting the data delivery delays for the different SG applications.

In order to reduce the amount of data sent over the network, and thus reduce energy consumption, data prediction is another potent solution of data reduction. It consists on predicting the values sensed by sensor nodes within certain error threshold, and resides both at the sensors and at the sink. The raw data is sent only if the desired accuracy is not satisfied, thereby reducing data transmission. We focus on time series estimation with Least Mean Square (LMS) for data prediction in WSN, in a Smart Grid context, where several applications with different data types and Quality of Service (QoS) requirements will exist on the same network. LMS proved its simplicity and robustness for a wide variety of applications, but the parameters selection (step size and filter length) can directly affect its global performance, choosing the right ones is then crucial. Having no clear and robust method on how to optimize these parameters for a variety of applications, we propose in [44] a modification of the original LMS that consists of training the filter for a certain time with the data itself in order to customize the aforementioned parameters. We consider different types of real data traces for the photo voltaic cells monitoring. Our simulation results provide a better data prediction while minimizing the mean square error compared to an existing solution in literature.

All these solutions have also been detailed in [12].

7.5. Connected Cars

Participants: Nathalie Mitton, Valeria Loscri, Joao Batista Pinto Neto.

7.5.1. Geolocalisation

Connected car technology promises to drastically reduce the number of accidents involving vehicles. Nevertheless, this technology requires the vehicle precise location to work. The adoption of Global Positioning System (GPS) as a navigation device imposes limitations to geolocation information under non-line-of-sight conditions. [22] introduces the Time Series Dead Reckoning System (TedriS) as a solution for dead reckoning navigation when the GPS fails. TedriS uses Time Series Regression Models (TSRM) and the data from the rear wheel speed sensor of the vehicle to estimate the absolute position. The process to estimate the position is carried out in two phases: training and predicting. In the training phase, a novel technique applies TSRM and stores the relationship between the GPS and the rear wheel speed data; then in the predicting phase, this relationship is used. We analyze TedriS using traces collected at the campus of Federal University of Rio de Janeiro (UFRJ), Brazil, and with indoor experiments with a robot. Results show an accuracy compatible with dead-reckoning navigation state-of-art systems.

7.5.2. Data forwarding

Intelligent inter-vehicle communication is a key research field in the context of vehicular networks that applies in real-life applications (e.g., management of accidents, intelligent fuel consumption, smart traffic jams, etc.). Considering different roles of nodes based on their "social aptitude" to relay information could provide a social component in the vehicular structure that can be useful in getting a clear prediction of the topological evolution in time and space proving to be very effective in managing intelligent data forwarding. In [36], we characterize a vehicular network as a graph using the link layer connectivity level and we classify nodes on the basis of specific attributes characterizing their "social aptitude" to forward data. Two forwarding approaches are presented, based on different socialites that allow to (i) select the most social node (i.e., a social hub) or (ii) choose among various social nodes.

7.5.3. Internet of vehicles

Internet, in its most recent evolution, is going to be the playground where a multitude of heterogeneous interconnected "things" autonomously exchange information to accomplish some tasks or to provide a service. Recently, the idea of giving to those smart devices the capability to organize themselves according to a social structure, gave birth to the so-called paradigm of the Social Internet of Things. The expected benefits of SIoT range from the enhanced effectiveness, scalability and speed of the navigability of the network of interconnected objects, to the provision of a level of trustworthiness that can be established by averaging the social relationships among things that are "friends". Bearing in mind the beneficial effects of social components in IoT, we consider a social structure in a vehicular context i.e., Social Internet of Vehicles (SIoV). In SIoV, smart vehicles build social relationships with other social objects they might come into contact, with the intent of creating an overlay social network to be exploited for information search and dissemination for vehicular applications. In [43], we aim to investigate the social behavior of vehicles in SIoV and how it is affected by mobility patterns. Specifically, through the analysis of simulated traffic traces, we distinguish friendly and acquaintance vehicles based on the encounter time and connection maintenance.

7.6. Robots and drones

Participants: Nathalie Mitton, Valeria Loscri, Farouk Mezghani, Anjalalaina Jean Cristanel Razafimandimby.

Internet of Robotic Things (IoRT) is a new concept introduced for the first time by ABI Research. Unlike the Internet of Things (IoT), IoRT provides an active sensorization and is considered as the new evolution of IoT. In this context, we propose a Neuro-Dominating Set algorithm (NDS) [21] to efficiently deploy a team of mobile wireless robots in an IoRT scenario, in order to reach a desired inter-robot distance, while maintaining global connectivity in the whole network. We use the term Neuro-Dominating Set to describe our approach, since it is inspired by both neural network and dominating set principles. With NDS algorithm, a robot adopts different behaviors according whether it is a dominating or a dominated robot. Our main goal is to show and demonstrate the beneficial effect of using different behaviors in the IoRT concept. The obtained results show that the proposed method outperforms an existing related technique (i.e., the Virtual Angular Force approach) and the neural network based approach presented in our previous work. As an objective, we aim to decrease the overall traveled distance and keep a low energy consumption level, while maintaining network connectivity and an acceptable convergence time.

Routing a fleet of robots in a known surface is a complex problem. It consists in the determination of the exact trajectory each robot has to follow to collect information. This is what we propose in [32] with the objective is to maximize the exploration of the given surface. To ensure that the robots can execute the mission in a collaborative manner, connectivity constraints are considered. These constraints guarantee that robots can communicate among each other and share the collected information. Moreover, the trajectories of the robots need to respect autonomy constraints.

When a disaster strikes, the telecommunications infrastructure gets damaged making rescue operations more challenging. Connecting first responders through flying base stations (i.e. drone mounted LTE (Long-Term Evolution) femtocell base station) presents a promising alternative to support infrastructure failure during disasters. The drone can travel the area and communicate with ground mobile devices, such as smartphones, and serves as flying data link to share information between survivors and rescuers. Problem statement. We would like to submit the following open problem to the community. Given the position of the ground mobile devices to serve, the problem presented here is about the dynamic drone path planning. As the drone autonomy is very limited and due to the high cost of drone mounted base station, the goal of this problem is to determine the best energy-efficient and minimum-time path to travel the area as fast as possible while still remaining in range of each survivor long enough to assure full servicing. This is the problem stated in [31].

7.7. MAC mechanisms

Participant: Nathalie Mitton.

In the era of the Internet of Things (IoT), the number of connected devices is growing dramatically. Often, connected objects use Industrial, Scientific and Medical (ISM) radio bands for communication. These kinds of bands are available without license, which facilitates development and implementation of new connected objects. However, it also leads to an increased level of interference in these bands. Interferences not only negatively affect the Quality of Service, but also cause energy losses, which is especially unfavorable for the energy constrained Wireless Sensor Networks (WSN). In [25], we develop an explicit formula of outage probability in a distributed wireless sensor network (WSN), assuming the MAC layer protocol being a slotted-ALOHA. And adopting a Markovian approach, we develop a model that analyses the performance of the slotted-ALOHA in order to improve these performances, in particular, by adding a preliminary stage of channel reservation, we show that this modification is important to have a high performance distributed wireless sensor network.

Several wild animal species are endangered by poaching. As a solution, deploying wireless sensors on animals able to send regular messages and also alert messages has been envisaged recently by several authorities and foundations. In that context, we have proposed WildMAC [35], a multichannel, multihop wireless communication protocol for these specific wireless sensor networks that have to collect data from unknown large areas with different QoS requirements. WildMAC is a TDMA based MAC protocol that leverages long range communication properties to propose an efficient data collection mean. Its performance evaluation shows it meets QoS requirements. To size the different parameters of WildMAC, we relied on the results of the study of [25].

7.8. RFID

Participants: Nathalie Mitton, Abdoul Aziz Mbacke, Ibrahim Amadou.

While RFID technology is gaining increased attention from industrial community deploying different RFID-based applications, it still suffers from reading collisions. As such, many proposals were made by the scientific community to try and alleviate that issue using different techniques either centralized or distributed, monochannel or multichannels, TDMA or CSMA. However, the wide range of solutions and their diversity make it hard to have a clear and fair overview of the different works. [18] surveys the most relevant and recent known state-of-the-art anti-collision for RFID protocols. It provides a classification and performance evaluation taking into consideration different criteria as well as a guide to choose the best protocol for given applications depending on their constraints or requirements but also in regard to their deployment environments.

Among all these approaches, [29], [28] propose new reader anti-collision schemes and data-priority aware data collection in a multi-hop RFID data collection protocol. [28] examines the implementation of two applications: for industrial IoT and for smart cities, respectively. Both applications, in regards to their requirements and configuration, challenge the operation of a RFID sensing solution combined with a dynamic wireless data gathering over multihops. They require the use of both mobile and fixed readers to cover the extent of deployment area and a quick retrieval of tag information. We propose a distributed crosslayer solution for improving the efficiency of the RFID system in terms of collision and throughput but also its proficiency in terms of tag information routing towards one or multiple sinks. Simulation results show that we can achieve high level of throughput while maintaining a low level of collision and a fairness of reader medium access above 95% in situations where readers can be fix and mobile, while tag information is routed with a data rate of 97% at worst and reliable delays for considered applications. [29] proposes cross-layer solutions meant for both scheduling of readers' activity to avoid collisions, and a multihop routing towards base stations, to gather read tag data. This routing is performed with a data priority aware mechanism allowing end-to-end delay reduction of urgent data packets delivery up to 13% faster compared to standard ones. Using fuzzy logic, we combine several observed metrics to reduce the load of forwarding nodes and improve latency as well as data rate. We validate our proposal running simulations on industrial and urban scenarios.

All these solutions have also been detailed in [11].

7.9. Smart Cities

Smart cities are a key factor in the consumption of materials and resources. As populations grow and resources become scarcer, the efficient usage of these limited goods becomes more important. Building on and integrating with a huge amount of data, the cities of the future are becoming a realization today. There are millions of sensors in place already, monitoring various things in metropolises. In the near future, these sensors will multiply until they can monitor everything from streetlights and trashcans to road conditions and energy consumption. In this context, effective strategies or solutions for refining data sets can play a key role. In [37], we propose a scheme in which passive RFID is shown as an interesting alternative and complement to WSN to alleviate the cost of some Smart City applications.

Also, in Smart Cities, crowd sensing may help to identify the current speed for each street, the congested areas, etc. In this context, map matching techniques are required to map a sequence of GPS waypoints into a set of streets on a common map. Unfortunately, most map matching approaches are probabilistic. In [41], we propose rather an unambiguous algorithm, able to identify all the possible paths that match a given sequence of waypoints. We need an unambiguous identification for each waypoints set. For instance, the actual speed should be assigned to the correct set of streets, without error. To identify all the possible streets, we construct the set of candidates iteratively. We identify all the edge candidates around each waypoint, and reconstruct all the possible sub-routes that connect them. We then verify a set of constraints, to eliminate impossible routes. The road segments common to all computed routes form an unambiguous match. We evaluate the matching ratio of our technique on real city maps (London, Paris and Luxembourg). We also validate our approach with a real GPS trace in Seattle.

In parallel, we proposed a MOOC in the framework of the IPL CityLab project (See Section 9.2.1), whose working documents are available online [48].

GAIA Team

7. New Results

7.1. Regular (differential) chains

[17] provides new equivalence theorems for regular chains and regular differential chains, which are generalizations of Ritt's characteristic sets. These theorems focus on regularity properties of elements of residue class rings defined by these chains, which are revealed by resultant computations. New corollaries to these theorems have quite simple formulations.

[30] contains a description of the management of the parameters in the`Maple DifferentialAlgebra package` and, in particular, in the RosenfeldGroebner function.

7.2. Systems of integro-differential equations

[28], [29] present a proof of concept for symbolic and numeric methods dedicated to the parameter estimation problem for models formulated by means of nonlinear integro-differential equations. In particular, we address the computation of the model input-output equation and the numerical integration of integro-differential systems (the BLINEIDE library).

7.3. Certified non-conservative tests for the structural stability of discrete multidimensional systems

In collaboration with Fabrice Rouillier (Inria Paris, Ouragan), in [18], we propose a new approach for testing the stability of nD systems. We first show that the standard characterization of the structural stability of a multivariate rational transfer function (namely, the denominator of the transfer function does not have solutions in the unit polydisc of $\mathbb{C}^n$) is equivalent to the fact that a certain system of polynomials does not have real solutions. We then use state-of-the-art computer algebra algorithms to check this last condition, and thus the structural stability of multidimensional systems. Our results have been implemented in a `Maple` prototype.

7.4. Using symbolic computation to solve algebraic Riccati equations arising in invariant filtering

In this joint work with Axel Barrau from Safran Tech [23], we propose a new step in the development of invariant observers. In the past, this theory led to impressive simplifications of the error equations encountered in estimation problems, especially those related to navigation. This was used to reduce computation load or derive new theoretical properties. Here, we leverage this advantage to obtain closed-form solutions of the underlying algebraic Riccati equations through advanced symbolic computation methods.

7.5. Parametric sub-optimal H_∞ controllers for an optro-mechanical system

In collaboration with *Safran Electronics & Defense*, in [15], we studied the robust stabilization of the line of sight of a stabilized mirror system. This system can be modeled by a single-input single- output time-delay system. Due to large model uncertainties, non-parametric methods are usually too conservative. Hence, we consider here unfixed model parameters. Using an additive decomposition, we show how to compute parametric H_∞ controllers of the time-delay model. Such a symbolic approach is interesting in the context of adaptive control and is illustrated throughout a simulation with an ideal parameter estimator.

7.6. A symbolic approach for signal demodulation and application to gearbox vibration analysis

This work is made in collaboration with Axel Barrau and Elisa Hubert (Safran Tech), and Roudy Dagher (Research Engineer, Inria Chile). The problem under study, which reduces to a certain signal factorization problem, was shown by Barrau et. al. to be equivalent to a Frobenius norm minimization problem. Starting from this optimization problem, we investigate the use of computer algebra methods to compute explicit solutions for the original problem. Along the way, we exhibit interesting algebraic and geometric properties of the underlying polynomial system. A paper is currently in development to summarize these results.

7.7. Curve analysis for the stability of time-delay systems

This work aims to design a new symbolic-numerical Puiseux-free approach for the study of the stability of differential time-delay systems. The idea behind is to replace the costly computations of Puiseux developpements around the *critical pairs* of the characteristic function by the numerical analysis of the branches of a well chosen 3D curve. The preliminary results show that this approach is easier to implement and turns out to be more efficient in practice. This ongoing work will be the subject of a future publication.

INOCS Project-Team

7. New Results

7.1. Large scale complex structure optimization

Formulation and algorithms for last-mile delivery systems:

E-commerce is a thriving market around the world and suits very well the busy lifestyle of today's customers and this growing e-commerce poses a huge challenge for transportation companies, especially in the last mile delivery. We addressed first a fleet composition problem for last-mile delivery service. This problem occurs at a tactical level when the composition of the fleet has to be decided in advance. It is the case for companies that offer last-mile delivery service. Most of them subcontract the transportation part to local carriers and have to decide the day before which vehicles will be needed to cover a partially known demand. We assumed that the distribution area is divided into a limited number of delivery zones and the time horizon into time-slots. The demand is characterized by packages to be transported from pick-up zones to delivery zones given a delivery time slot. First, we introduced an integer programming model which aims to minimize the total delivery cost while ensuring that the demand is covered, the capacity of each vehicle is not violated, the working time for each period is not exceeded and the total working of each delivery respects the social regulations. Then we present a column-generation based approach, which is able to solve real-life instances in reasonable CPU times [33], [32]. Nowadays, the most common last mile delivery service is home delivery. Besides home delivery, companies like Amazon and Fedex, develop locker delivery. When customers shop online, they can choose a nearby locker as a pickup location. In the past years, a new concept called trunk delivery, has been proposed. Here, customers' orders can be delivered to the trunk of their cars. We jointly considered all these delivery possibilities in the same last-mile system and studied the case where the fleet is limited to a single vehicle. We proposed different formulations for the rising optimization problem. We developed problem defined cuts in order to strengthen the formulations and be able to tackle real-size instances. Last we designed and implemented a branch-and-cut algorithm [55], [53].

Large neighborhood algorithm for multi-commodity vehicle routing problem: When delivering fresh fruits and vegetables to catering the multi commodity aspect should be taken into account and deliveries to customers are not made in once, but each commodity can be delivered by a different vehicle as long as the total demand of that commodity is delivered. The problem that arises is the commodity constrained split delivery vehicle routing problem (C-SDVRP). We propose a heuristic based on the adaptive large neighborhood search (ALNS) to efficiently solve medium and large sized instances of the C-SDVRP. We take into account the distinctive features of the C-SDVRP and adapt several local search moves to improve a solution. Moreover, a mathematical programming based operator (MPO) that reassigns commodities to routes is used to improve a new global best solution. Computational experiments have been performed on benchmark instances from the literature. The results assess the efficiency of the algorithm, which can provide a large number of new best-known solutions in short computational times [50].

A matheuristic for the packaging and shipping problem: E-commerce has been continuously growing in the last years to a primary retail market. Recently in France, the threshold of 1 billion of online transactions was overcome. Due to a high demand fluctuation of e-commerce, the workforce sizing for the logistic chain is a challenging problem. Companies have to develop good strategies to have a sustainable workforce size while guaranteeing a high-level service. In this work, we consider the management of the workforce for a warehouse of an e-commerce company. Specifically, we address issues as i) How the workforce at the warehouse can be determined; ii) What is the daily operational production planning; iii) How the demand peaks can be smoothed, and the production maintained ideally constant over the time horizon. To provide answers to these issues, we introduce the Packaging and Shipping Problem (PSP). The PSP looks for a solution approach that jointly determines the workforce over a multi-period horizon and daily operational plans while minimizing the total logistics cost. We consider two strategies that aim to enhance the flexibility of the process and the efficiency

of resources use: reassignment and postponement. To tackle the Packaging and Shipping Problem we propose a model, and a three-phase matheuristic. This heuristic is proved to be competitive with respect to the direct solution of the model with a commercial solver on real-life based instances [18].

Heuristic and column generation approaches for the joint order batching and picker routing problem: Picking is the process of retrieving products from the inventory. It is mostly done manually by dedicated employees called pickers and is considered the most expensive of warehouse operations. To reduce the picking cost, customer orders can be grouped into batches that are then collected by traveling the shortest possible distance. We proposed an industrial case study for the HappyChic company where the warehouse has an acyclic layout: pickers are not allowed to backtrack. We developed a two-phase heuristic approach to solve this industrial case [59]. Moreover, we proposed an exponential linear programming formulation to tackle the joint order batching and picker routing problem. Variables, or columns, are related to the picking routes in the warehouse. Computing such routes is in general an intractable routing problem and relates to the well known traveling salesman problem (TSP). Nonetheless, the rectangular warehouse's layouts can be used to efficiently solve the corresponding TSP. Experimented on a publicly available benchmark, the algorithm proves to be very effective. It improves many of the best known solutions and provides very strong lower bounds. This approach is also applied to the HappyChic industrial case to demonstrate its interest for this field of application [41].

Distribution network configuration problems: A distribution network is a system aiming to transfer a certain type of resource from feeders to customers. Feeders are producers of a resource and customers have a certain demand in this resource that must be satisfied. Distribution networks can be represented on graphs and be subject to constraints that limit the number of intermediate nodes between some elements of the network (hop constraints) because of physical constraints. We used layered graphs for hop constrained problems to build extended formulations [21]. Preprocessing techniques allowed to reduce the size of the layered graphs used. The model was studied on the hop-constrained minimum margin problem in an electricity network. This problem consists of designing a connected electricity distribution network, and to assign customers to electricity feeders at a maximum number of hops H so as to maximize the minimum capacity margin over the feeders to avoid an overload for any feeder. A related theoretical work considers a very special case of hop constrained network design, namely the 2 edge-disjoint 3-paths polyhedron [15].

Switched Ethernet network design problems: We studied models arising in the design of switched Ethernet networks implementing the Multiple Spanning Tree Protocol [23]. In these problems, multiple spanning trees have to be established in a network to route demands partitioned into virtual local access networks. Different mixed-integer formulations for the problem have been proposed and compared, both theoretically and computationally.

Delay management in public transportation: The Delay Management Problem arises in Public Transportation networks, and is characterized by the necessity of connections between different vehicles. The attractiveness of Public Transportation networks is strongly related to the reliability of connections, which can be missed when delays or other unpredictable events occur. Given a single initial delay at one node of the network, the Delay Management Problem is to determine which vehicles have to wait for the delayed ones, with the aim of minimizing the dissatisfaction of the passengers. We derived strengthened mixed integer linear programming formulations and new families of valid inequalities for that problem. The implementation of branch-and-cut methods and tests on a benchmark of instances taken from real networks show the potential of the proposed formulations and cuts [20].

Discrete ordered median problem: The discrete ordered median problem consists in locating p facilities in order to minimize an ordered weighted sum of distances between clients and closest open facility. We formulate this problem as a set partitioning problem using an exponential number of variables. Each variable corresponds to a set of demand points allocated to the same facility with the information of the sorting position of their corresponding costs. We develop a column generation approach to solve the continuous relaxation of this model. Then, we apply a branch-price-and-cut algorithm to solve small to large sized instances of DOMP in competitive computational time [62].

Genome wide association studies: We studied the Polymorphic Alu Insertion Recognition Problem (PAIRP). Alu (Arthrobacter luteus) forms a major component of repetitive DNA and are frequently encountered during the genotyping of individuals. The basic approach to find Alus consists of (i) aligning sequence reads from a set of individual(s) with respect to a reference genome and (ii) comparing the possible Alu insertion induced by the alignment with the Alu insertions positions already known for the reference genome. The sequence genome of the reference individual is known and will be highly similar, but not identical, to the genome of the individual(s) being sequenced. Hence, at some locations they will diverge. Some of this divergence is due to the insertion of Alu polymorphisms. Detecting Alus has a central role in the field of Genetic Wide Association Studies because basic elements are a common source of mutation in humans. We investigated the PAIRP relationship with the the Clique Partitioning of Interval Graphs (CPIG). Our results [12], [26] provide insights of the complexity of the problem, a characterization of its combinatorial structure and an exact approach based on Integer Linear Programming to exactly solve the correspond instances.

A branch-and-cut algorithm for the maximum k-balanced subgraph of a signed graph: A signed graph is k-balanced if its vertex set can be partitioned into at most k sets in such a way that positive edges are found only within the sets and negative edges go between sets. The maximum k-balanced subgraph problem is the problem of finding a subgraph of G that is k-balanced and maximum according to the number of vertices. This problem has applications in clustering problems appearing in collaborative vs conflicting environments. We provide a representatives formulation for the problem and present a partial description of the associated polytope, including the introduction of strengthening families of valid inequalities. A branch-and-cut algorithm is described for finding an optimal solution to the problem. An ILS metaheuristic is implemented for providing primal bounds for this exact method and a branching rule strategy is proposed for the representatives formulation. Computational experiments, carried out over a set of random instances and on a set of instances from an application, show the effectiveness of the valid inequalities and strategies adopted in this work [22].

Feature selection in support vector machine: This work focuses on support vector machine (SVM) with feature selection. A MILP formulation is proposed for the problem. The choice of suitable features to construct the separating hyperplanes has been modelled in this formulation by including a budget constraint that sets in advance a limit on the number of features to be used in the classification process. We propose both an exact and a heuristic procedure to solve this formulation in an efficient way. Finally, the validation of the model is done by checking it with some well-known data sets and comparing it with classical classification methods [25].

7.2. Bilevel Programming

Pricing problems in energy management: Power systems face higher flexibility requirements from generation to consumption due to the increasing penetration of non-controllable distributed renewable energy. In this context, demand side management aims at reducing excessive load fluctuation and match the price of energy to their real cost for the grid. Pricing models for demand side management methods are traditionally used to control electricity demand. First, we proposed bilevel pricing models to explore the relationship between energy suppliers and customers who are connected to a smart grid. The smart grid technology allows customers to keep track of hourly prices and shift their demand accordingly, and allows the provider to observe the actual demand response to its pricing strategy. Moreover, we assumed that the smart grid optimizes the usage of a renewable energy generation source and a storage capacity. Results over a rolling horizon were obtained [14], [28], [36]. Next, we considered four types of actors: furnishers sell electricity, local agents trade and consume energy, aggregators trade energy and provide energy to end-users, who consume it. This gives rise to three levels of optimization. The interaction between aggregators and their end-users is modelled with a bilevel program, and so is the interaction between furnishers, and local agents and aggregators. Since solving bilevel programs is difficult in itself, solving trilevel programs requires particular care. We proposed three possible approaches, two of them relying on a characterization of the intermediary optimization level [35]. Finally, Time and-Level-of-Use is a recently proposed energy pricing scheme, designed for the residential

sector and providing suppliers with robust guarantee on the consumption. We formulate the supplier decision as a bilevel, bi-objective problem optimizing for both financial loss and guarantee. A decomposition method is proposed, related to the optimal value transformation. It allows for the computation of an exact solution by finding possible Pareto optimal candidate solutions and then eliminating dominated ones. Numerical results on experimental residential power consumption data show the method effectively finds the optimal candidate solutions while optimizing costs only or incorporating risk aversion at the lower-level [37].

Unit commitment under market equilibrium constraints: Traditional (deterministic) models for the Unit Commitment problem (UC) assume that the net demand for each period is perfectly known in advance, or in more recent and more realistic approaches, that a set of possible demand scenarios is known (leading to stochastic or robust optimization problems). However, in practice, the demand is dictated by the amounts that can be sold by the producer at given prices on the day-ahead market. We modeled and solved the UC problem with a second level of decisions ensuring that the produced quantities are cleared at market equilibrium. In its simplest form, we are faced to a bilevel optimization problem where the first level is a MIP and the second level linear. As a first approach to the problem, we assumed that demand curves and offers of competitors in the market are known to the operator. Following the classical approach for these models, we turned the problem into a single-level program by rewriting and linearizing the first-order optimality conditions of the second level. In recent work, this approach was extended to include network capacities effects and a decoupling of prices in different zones [45], [46], [47], [48].

Market regulation: We proposed a bilevel programming model to study a problem of market regulation through government intervention. One of the main characteristics of the problem is that the government monopolizes the raw material in one industry, and competes in another industry with private firms for the production of commodities. Under this scheme, the government controls a state-owned firm to balance the market; that is, to minimize the difference between the produced and demanded commodities. On the other hand, a regulatory organism that coordinates private firms aims to maximize the total profit by deciding the amount of raw material bought from the state-owned firm. Two equivalent single-level reformulations are proposed to solve the problem. Additionally, three heuristic algorithms are designed to obtain good-quality solutions with low computational effort. Extensive computational experimentation is carried out to measure the efficiency of the proposed solution methodologies. A case study based on the Mexican petrochemical industry is presented. Additional instances generated from the case study are considered to validate the robustness of the proposed heuristic algorithms [66].

Rank pricing: One of the main concerns in management and economic planning is to sell the right product to the right customer for the right price. Companies in retail and manufacturing employ pricing strategies to maximize their revenues. The Rank Pricing Problem considers a unit-demand model with unlimited supply and uniform budgets in which customers have a rank-buying behavior. Under these assumptions, the problem is first analyzed from the perspective of bilevel pricing models and formulated as a non linear bilevel program with multiple independent followers. We also present a direct non linear single level formulation. Two different linearizations of the models are carried out and two families of valid inequalities are obtained which, embedded in the formulations by implementing a branch-and-cut algorithm, allow us to tighten the upper bound given by the linear relaxation of the models. We show the efficiency of the formulations, the branch-and-cut algorithms and some preprocessing through extensive computational experiments [16].

Bilevel minimum spanning tree problem: Consider a graph whose edge set is partitioned into a set of red edges and a set of blue edges, and assume that red edges are weighted and contain a spanning tree of G. Then, the Bilevel Minimum Spanning Tree Problem (BMSTP) consists in pricing (i.e., weighting) the blue edges in such a way that the total weight of the blue edges selected in a minimum spanning tree of the resulting graph is maximized. We propose different mathematical formulations for the BMSTP based on the properties of the Minimum Spanning Tree Problem and the bilevel optimization. We establish a theoretical and empirical comparison between these new formulations and we also provide reinforcements that together with a proper formulation are able to solve medium to big size instances [65].

Bilevel programming models for location problems: First, we addressed a multi-product location problem in which a retail firm has several malls with a known location. A particular product comes in p types. Each mall has a limited capacity for products to be sold at that location, so the firm has to choose what products to sold at what mall. Furthermore, the firm can apply discrete levels of discount on the products/ The objective of the firm is to find what products to sell at which mall, with what level of discount, so that its profit is maximized. Consumers are located in points of the region. Each consumer has a different set of acceptable products, and will purchase one of these, or none if it is not convenient for her. Consumers maximize their utility. The agents (firm and consumers) play a Stackelberg game, in which the firm is the leader and the customers the follower. Once the firm decides the products to sell at each mall and the possible discounts, consumers purchase (or not) one of their acceptable products wherever their utility is maximized. We model the problem using bilevel formulations, which are compared on known instances from the literature [43]. Second we studied a location problem of controversial facilities. On the one hand, a leader chooses among a number of fixed potential locations which ones to establish. On the second hand, one or several followers who, once the leader location facilities have been set, choose their location points in a continuous framework. The leader's goal is to maximize some proxy to the weighted distance to the follower's location points, while the follower(s) aim is to locate his location points as close as possible to the leader ones. We develop the bilevel location model for one follower and for any polyhedral distance, and we extend it for several followers and any so-called p-norm. We prove the NP-hardness of the problem and propose different mixed integer linear programming formulations. Moreover, we develop alternative Benders decomposition algorithms for the problem. Finally, we report some computational results comparing the formulations and the Benders decompositions on a set of instances [63].

Stackelberg games: First we analyzed general Stackelberg games (SGs) and Stackelberg security games (SSGs). SGs are hierarchical adversarial games where players select actions or strategies to optimize their payoffs in a sequential manner. SSGs are a type of SGs that arise in security applications, where the strategies of the player that acts first consist in protecting subsets of targets and the strategies of the followers consist in attacking one of the targets. We review existing mixed integer optimization formulations in both the general and the security setting and present new formulations for the the second one. We compare the SG formulations and the SSG formulations both from a theoretical and a computational point of view. We identify which formulations provide tighter linear relaxations and show that the strongest formulation for the security version is ideal in the case of one single attacker. Our computational experiments show that the new formulations can be solved in shorter times [61].

Second, we formulate a Stackelberg Security game that coordinates resources in a border patrol problem. In this security domain, resources from different precincts have to be paired to conduct patrols in the border due to logistic constraints. Given this structure the set of pure defender strategies is of exponential size. We describe the set of mixed strategies using a polynomial number of variables but exponentially many constraints that come from the matching polytope. We then include this description in a mixed integer formulation to compute the Strong Stackelberg Equilibrium efficiently with a branch and cut scheme. Since the optimal patrol solution is a probability distribution over the set of exponential size, we also introduce an efficient sampling method that can be used to deploy the security resources every shift. Our computational results evaluate the efficiency of the branch and cut scheme developed and the accuracy of the sampling method. We show the applicability of the methodology by solving a real world border patrol problem [58].

7.3. Robust/Stochastic programming

Locating stations in a one-way electric car sharing system under demand uncertainty: We focused in [60] on a problem of locating recharging stations in one-way station based electric car sharing systems which operate under demand uncertainty. We modeled this problem as a mixed integer stochastic program and develop a Benders decomposition algorithm based on this formulation. We integrated a stabilization procedure to our algorithm and conduct a large-scale experimental study on our methods. To conduct the computational experiments, we developed a demand forecasting method allowing to generate many demand scenarios. The method was applied to real data from Manhattan taxi trips.

Integrated shift scheduling and load assignment optimization for attended home delivery: We studied an integrated shift scheduling and load assignment optimization problem for attended home delivery. The proposed approach is divided into two phases, each one corresponding to a different planning level: tactical and operational. In the tactical planning, a daily master plan is generated for each courier. This master plan defines the working shifts, the origin-destination pairs to visit, and the number of client requests to serve. In the operational planning, delivery orders are allocated to couriers in real-time. The stochastic and dynamic nature of client orders is included in the tactical and operational decision levels, respectively. For the tactical level, we developed and implemented a multi-cut L-shaped algorithm. Experimental results demonstrate that our approach provides robust tactical solutions that easily accommodate to fluctuations in client orders, preventing additional costs related to the underutilization of couriers and to the use of external couriers to satisfy all delivery requests, when compared to an approach using the mean demand value. Moreover, these results also indicate that the failure to incorporate robust tactical solutions in the operational planning results in infeasible operational plans that are inadmissible regarding the couriers' working time (e.g., minimum and maximum number of working hours) and work regulations (e.g., allocation of consecutive working hours to the couriers).

Bookings in the European gas market: Characterization of feasibility and computational complexity results: As a consequence of the liberalisation of the European gas market in the last decades, gas trading and transport have been decoupled. At the core of this decoupling are so-called bookings and nominations. Bookings are special long-term capacity right contracts that guarantee that a specified amount of gas can be supplied or withdrawn at certain entry or exit nodes of the network. These supplies and withdrawals are nominated at the day-ahead. These bookings then need to be feasible, i.e., every nomination that complies with the given bookings can be transported. While checking the feasibility of a nomination can typically be done by solving a mixed-integer nonlinear feasibility problem, the verification of feasibility of a set of bookings is much harder. We consider the question of how to verify the feasibility of given bookings for a number of special cases. For our physics model we impose a steady-state potential-based flow model and disregard controllable network elements. We derive a characterization of feasible bookings, which is then used to show that the problem is in coNP for the general case but can be solved in polynomial time for linear potential-based flow models. Moreover, we present a dynamic programming approach for deciding the feasibility of a booking in tree-shaped networks even for nonlinear flow models [56].

LINKS Project-Team

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Data Integration

The PhD project of Lozano on relational to RDF data integration is progressing under the direction of Boneva, and Staworko. At **AMW** [9] they studied the *relational to RDF data exchange problem*. They focus in particular on a preliminary analysis of the consistency problem for relational to RDF data exchange with target ShEx schema.

7.1.2. Schema Validation

Shape Expression Language 2.0 (ShEx) is a language to describe the vocabulary and the structure of an RDF graph. It is base on the notion of shapes, a typing system supporting algebraic operations, recursive references to other shapes or Boolean combination.

In their **PODS** paper [7], Staworko studied the *containment problem* for ShEx (in cooperation with Wieczorek from Wrazlaw). Containement is a classical subject for schema-related issue in database theory. The authors proved that it is decidable for ShEx-schema, but with a untractable complexity (co-NEXP-hard). They also carefuly craft restriction of ShEx schema to design tractable-but-still-signifiant fragments.

7.2. Managing Dynamic Linked Data

7.2.1. Complex Event Processing

Complex event processing requires to answer queries on streams of complex events, i.e., nested words or equivalently linearizations of data trees, but also to produce dynamically evolving data structures as output.

The topic of the PhD project of M. Sakho supervised by Niehren and Boneva is to generalize algorithms for querying streams to hyperstreams. These are collections of linked streams as naturally produced as intermediate results of complex events processing. Hyperstreams are incomplete descriptions of relational structures, so they can queried similarly to incomplete databases, for which the notion of a certain query answer is most appropriate.

In a paper published at **RP** [13], they studied certain query answering for hyperstreams with simple events. Such hyperstreams can be identified with compressed string patterns. They proved that the certain query answering for regular queries on compressed string patterns is PSPACE-complete, independently of whether the finite automata defining the regular queries are assumed deterministic or not, and independently of whether compression is permitted or not. They also showed that the problem is in PTIME when restricted to *linear* string patterns (possibly with compression) and to deterministic finite automata.

In a paper published at **LATA** [6], the studied certain query answering on hyperstreams of complex events. Such hyperstreams can be modeled by compressed tree pattern with context variables. They showed that certain query answering for regular queries on compressed tree pattern with context variables is EXP-complete, independently of whether the tree automata defining the regular queries are assumed deterministic or not, and independently of whether compression is permitted or not. They also showed that the problem is in PTIME when restricted to *linear* tree patterns (possibly with compression) and to deterministic tree automata.

7.2.2. Transformations

In his PhD project – belonging to the ANR Colis– Gallot with his supervisors Salvati and Lemay presented higher order tree transducers which extend macro tree transducers. Moreover they obtained nice properties such as the closure of the transducers under composition. Algorithms to compute such compositions are proposed. Those algorithms perform partial evaluation and are guided by semantic interpretations over finite domains.

Another virtue of higher-order transducers is that their *linear* syntactic restriction make them equivalent to logically defined MSO transductions. One of the composition algorithm proposed preserves the linearity. Furthermore, we have also showed that we can decrease the order of linear transducer (i.e. the complexity of the functions it handles) when this one is larger than 4.

These results are unpublished paper for now.

7.3. Foundations of AI

Various problems of databases and knowledge bases are closely related to foundational problems in artificial intelligence, since they are rooted in logic or graph theory.

7.3.1. Knowledge Compilation

Many problems in Artificial Intelligence boil down to the exploration of the solution set (called the models) of logical formulas. Such an exploration can be finding one model of the formula, counting the number of models or enumerating them all. However, even for simple quantifier-free formulas, those explorations are known be untractable (NP-hard).

Knowledge compilation encompasses methods that aim to change the representation of the set of models in order to get tractable algorithms for (some of) those tasks. A big computational cost is paid during the compilation time but then replying to queries become tractable on the new representation. More generally, the core of Knowledge compilation is the study of the trade-off between the size of the representation and the easiness of queries. This subject is of interest for both Artificial Intelligence and Database communities.

At **STACS** [15], Capelli, in cooperation with Mengel from CRIL (Lens), studied knowledge compilation techiques for quantified Boolean formulas. Deciding the existence of models for such formulas is known to climb arbitrarly high the polynomial time hierarchy. The authors provide an efficient compilation procedure for formulas having a *bounded tree-width* generalizing results from SAT solving.

7.3.2. Aggregation and Enumeration for Graphs

Aggregation and enumeration are not relevant for answer sets of database queries but equally for any kinds of sets, most typically defined by combinatoric problems on graphs.

In a paper published at **ICALP** [8], Paperman proposed (in cooperation with Amarilli from Telecom Paristech) to study the problem of finding so called *topological sort* satisfying constraints provided by regular expressions. Searching topological sort happens typically in situations where an order is *uncertain*. For instance, in relational database where users provides a partial preference order, or in concurrent and distributed programming where some tasks can be executed in an arbitrary order. A classical task in *preferential query answering* is to find a topological sort satisfying some global constrained. Typically, to find a total order satisfying all (or most) of the customers. The paper provides and proves sufficient conditions on the *shape of the constraints* to make the problem tractable (P-time) as well as sufficient condition to make the problem NP-hard. They also prove a complete dichotomy for an adapted and well chosen version of the constrained topological sort problem.

In an article in **JCSS** [2], Capelli (with Bergougnoux and Kanté from Bordeaux and Clérmont-Ferrand) propose an algorithm for counting the number of *transversal* in some *hypergraphs*. Here, a hypergraph is a collection of sets – called *hyperedges* over a *ground set* and a traversal is a subset intersecting all hyperedges. In full generality, counting the number of minimal traversals in a hypergraph is a hard problem: it is known to be $\#P$-complete. They proved that under the assumptions of β-acyclicity, it is possible to count all the minimal traversals can be done in polynomial times.

LOKI Team

7. New Results

7.1. Introduction

According to our research program, in the next two to five years, we will study dynamics of interaction along three levels depending on interaction time scale and related user's perception and behavior: *Micro-dynamics*, *Meso-dynamics*, and *Macro-dynamics*. Considering phenomena that occur at each of these levels as well as their relationships will help us to acquire the necessary knowledge (Empowering Tools) and technological bricks (Interaction Machine) to reconcile the way interactive systems are designed and engineered with human abilities. Although our strategy is to investigate issues and address challenges for all of the three levels of dynamics, our immediate priority is to focus on micro-dynamics since it concerns very fundamental knowledge about interaction and relates to very low-level parts of interactive systems, which is likely to influence our future research and developments at other levels.

7.2. Micro-dynamics

Participants: Axel Antoine, Géry Casiez [correspondent], Sylvain Malacria, Mathieu Nancel, Thomas Pietrzak.

7.2.1. Latency & Transfer functions

End-to-end latency in interactive systems is detrimental to performance and usability, and comes from a combination of hardware and software delays. While these delays are steadily addressed by hardware and software improvements, it is at a decelerating pace. In parallel, short-term input prediction has recently shown promising results to compensate for latency, in both research and industry.

in the context of the collaborative Turbotouch project, we introduced a new prediction algorithm for direct touch devices based on (i) a state-of-the-art finite-time derivative estimator, (ii) a smoothing mechanism based on input speed, and (iii) a post-filtering of the prediction in two steps (see Figure 2 left). Using both a preexisting dataset of touch input as benchmark, and subjective data from a new user study, we showed that this new predictor outperforms those currently available in the literature and industry, based on metrics that model user-defined negative side-effects caused by input prediction. In particular, our predictor can predict up to 2 or 3 times further than existing techniques with minimal negative side-effects [23].

We also proposed a hybrid hardware and software input prediction technique specifically designed for partially compensating end-to-end latency in indirect pointing (see Figure 2 right). We combined a computer mouse with a high frequency accelerometer to predict the future location of the pointer using Euler based equations. Our prediction method results in more accurate prediction than previously introduced prediction algorithms for direct touch. A controlled experiment also revealed that it can improve target acquisition time in pointing tasks [15], [28].

Finally, on the topic of transfer functions we performed some preliminary analysis of the kinematics of a pointing task with varying linear velocity based transfer functions to assess how we use vision and haptics to plan and control our movement [25].

7.2.2. Understanding touch interaction

Atomic interactions in touch interfaces, like tap, drag, and flick, are well understood in terms of interaction design, but less is known about their physical performance characteristics. We conducted a study to gather baseline data about finger pitch and roll orientation during atomic touch input actions [21]. Our results showed differences in orientation and range for different fingers, hands, and actions: for a given hand, the little, ring and middle fingers are used in a similar manner, whereas the thumb uses different range of orientations. Additional analyses about how changing the angle of the tablet affects people's finger orientations suggest that

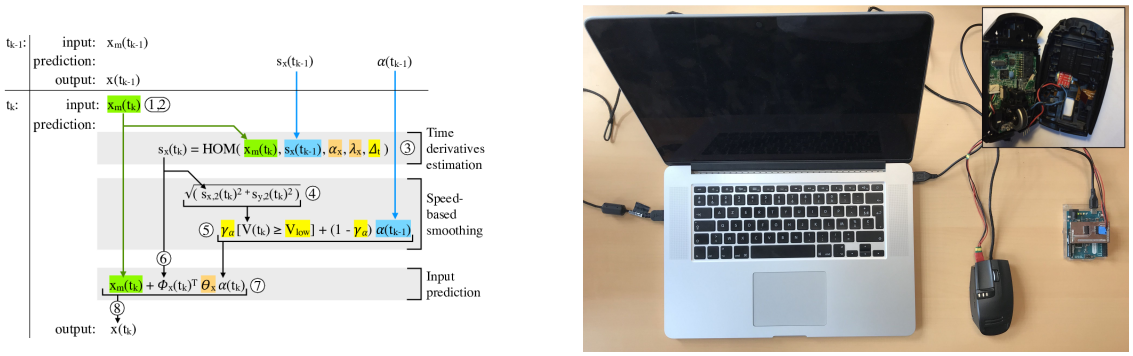

Figure 2. (left) General description of our real-time input prediction method, with step numbers. Input in green, previously computed variables in blue, general parameters in yellow, optimized parameters in orange. (right) Our hybrid setup for input prediction comprises a Logitech G9 Laser Mouse connected via USB to the host computer with the MPU-9250 chip embedded inside, which is itself connected to an Arduino board.

ranges of orientation tighten as the tablet pitch increases. This data provides designers and researchers with better understanding of what kind of interactions are possible in different settings (e. g., using the left or right hand), to design novel interaction techniques that use orientation as input (e. g., using finger tilt as an implicit mode), and to anticipate the feasibility of new sensing techniques (e. g., using fingerprints for identifying specific finger touches).

7.3. Meso-dynamics

Participants: Marc Baloup, Géry Casiez, Stéphane Huot, Edward Lank, Sylvain Malacria, Mathieu Nancel, Thomas Pietrzak [correspondent], Thibault Raffaillac, Marcelo Wanderley.

7.3.1. Improving interaction bandwidth and expressiveness

Despite the ubiquity of touch-based input and the availability of increasingly computationally powerful touchscreen devices, there has been comparatively little work on enhancing basic canonical gestures such as swipe-to-pan and pinch-to-zoom. We introduced transient pan and zoom, i. e., pan and zoom manipulation gestures that temporarily alter the view and can be rapidly undone [16]. Leveraging typical touchscreen support for additional contact points, we designed our transient gestures so that they co-exist with traditional pan and zoom interaction. In addition to reducing repetition in multi-level navigation, our transient pan-and-zoom also facilitates rapid movement between document states.

Image editing software feature various pixel selection tools based on geometrical (rectangle, ellipses, polygons) or semantical (magic wand, selection brushes) data from the image. They are efficient in many situations, but are limited when selecting bitmap representations of handwritten text for e. g., interpreting scanned historical documents that cannot be reliably analyzed by automatic OCR methods: strokes are thin, with many overlaps and brightness variations. We have designed a new selection tool dedicated to this purpose [27]: a cursor based brush selection tool with two additional degrees of freedom: brush size and brightness threshold. The brush cursor displays feedforward clues that indicates the user which pixels will be selected upon pressing the mouse button. This brush provides a fine grain control to the user over the selection.

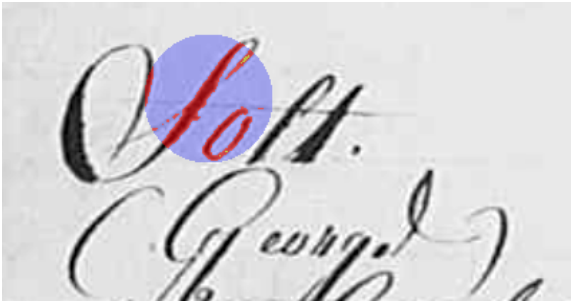

Figure 3. A four-dimensional selection brush for digitized handwritten documents. Red pixels will be selected, blue pixels will not.

7.3.2. Interacting with specific setups (Large-Displays, Virtual & Augmented Reality)

Large displays are becoming commonplace at work, at home, or in public areas. Handheld devices such as smartphones and smartwatches are ubiquitous, but little is known on regarding how these devices could be used to point at remote large displays. We conducted a survey on possession and use of smart devices, as well as a controlled experiment comparing seven distal pointing techniques on phone or watch, one- and two-handed, and using different input channels and mappings [26]. Our results favor using a smartphone as a trackpad, but also explore performance tradeoffs that can inform the choice and design of distal pointing techniques for different contexts of use.

In virtual reality environments, raycasting is the most common target pointing technique. However, performance on small and distant targets is impacted by the accuracy of the pointing device and the user's motor skills. Existing pointing facilitation techniques are currently only applied in the context of a virtual hand, i. e., for targets within reach. We studied how a user-controlled cursor could be added on the ray in order to enable target proximity-based pointing techniques –such as the Bubble Cursor– to be used for targets that are out of reach [17]. We conducted a study comparing several visual feedbacks for this technique (see Figure 4). Our results showed that simply highlighting the nearest target reduces the selection time by 14.8% and the error rate by 82.6% compared to standard Raycasting. For small targets, the selection time is reduced by 25.7% and the error rate by 90.8%.

Brain-Computer Interfaces (BCIs) enable users to interact with computers without any dedicated movement, bringing new hands-free interaction paradigms that could be beneficial in an Augmented Reality (AR) setup. We first tested the feasibility of using BCI in AR settings based on Optical See-Through Head-Mounted Displays (OST-HMDs) [12]. Experimental results showed that a BCI and an OST-HMD equipment (EEG headset and Hololens in our case) are well compatible and that small movements of the head can be tolerated when using the BCI. Then, we introduced a design space for command display strategies based on BCI in AR, when exploiting a famous brain pattern called Steady-State Visually Evoked Potential (SSVEP). Our design space relies on five dimensions concerning the visual layout of the BCI menu: orientation, frame-of-reference, anchorage, size and explicitness. We implemented various BCI-based display strategies and tested them within the context of mobile robot control in AR. Our findings were finally integrated within an operational prototype based on a real mobile robot that is controlled in AR using a BCI and a HoloLens headset. Taken together, our results (4 user studies) and our methodology could pave the way to future interaction schemes in Augmented Reality exploiting 3D User Interfaces based on brain activity and BCIs.

More generally, we also contributed to a reflexion on the complexity and scientific challenges associated to virtual and augmented realities [29] and the challenges to make virtual environments more closely related to the real world [30].

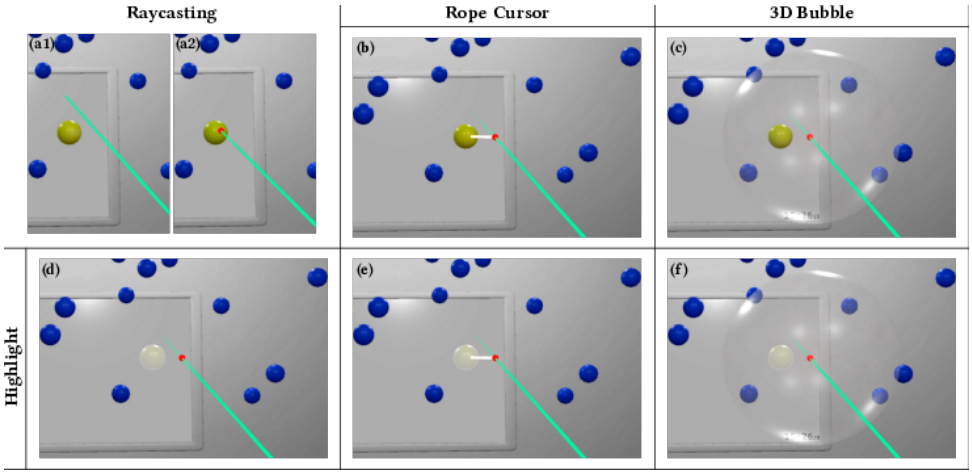

Figure 4. Visual feedback for RayCursor: (a1,a2) classical Raycasting; (b) Rope Cursor: a stroke between the closest target and the cursor; (c) 3D Bubble: a bubble centered on the cursor which contains the nearest target; (d) Highlighting on the nearest target; (e,f), highlight + rope and 3D Bubble.

7.3.3. Tools for prototyping and programming interaction

Touch interactions are now ubiquitous, but few tools are available to help designers quickly prototype touch interfaces and predict their performance. On one hand, for rapid prototyping, most applications only support visual design. On the other hand, for predictive modeling, tools such as CogTool generate performance predictions but do not represent touch actions natively and do not allow exploration of different usage contexts. To combine the benefits of rapid visual design tools with underlying predictive models, we developed the *Storyboard Empirical Modeling (StEM)* tool [20], [19] for exploring and predicting user performance with touch interfaces (see Figure 5). StEM provides performance models for mainstream touch actions, based on a large corpus of realistic data. We evaluated StEM in an experiment and compared its predictions to empirical times for several scenarios. The study showed that our predictions are accurate (within 7% of empirical values on average), and that StEM correctly predicted differences between alternative designs. Our tool provides new capabilities for exploring and predicting touch performance, even in the early stages of design.

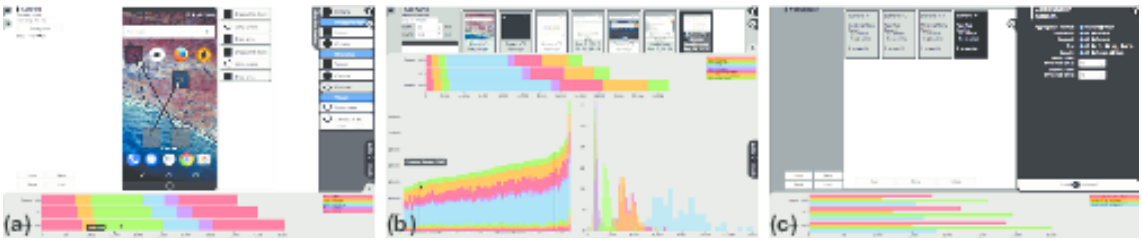

Figure 5. Storyboard Empirical Modeling (StEM): (a) users drag and drop actions onto a timeline to construct an interaction sequence; (b) users can visualize prediction times for a scenario composed of different screens; (c) users can compare scenarios, and filter the predictions according to contextual factors such as screen size or user's expertise.

Following our main objective of revisiting interactive system, we have also proposed two systems for defining and programming interactive behaviors and interactions.

Much progress has been made on interactive behavior development tools for expert programmers. However, less effort has been made in investigating how these tools support creative communities who typically struggle with technical development. This is the case, for instance, of media artists and composers working with interactive environments. To address this problem, we have introduces ZenStates [18], a new specification model for creative interactive environments that combines Hierarchical Finite-States Machines, expressions, off-the-shelf components called Tasks, and a global communication system called the Blackboard. We have implemented our model in a direct manipulation-based software interface and probed ZenStates' expressive power through 90 exploratory scenarios. we have also conducted a user study to investigate the understandability of ZenStates' model. Results support ZenStates viability, its expressiveness, and suggest that ZenStates is easier to understand –in terms of decision time and decision accuracy– compared to popular alternatives such as standard object-oriented programming and a data-flow visual language.

In a more general context, we have introduced a new GUI framework based on the *Entity-Component-System* model (ECS), where interactive elements (Entities) can acquire any data (Components) [24]. Behaviors are managed by continuously running processes (Systems) which select entities by the components they possess. This model facilitates the handling and reuse of behaviors. It allows to define the interaction modalities of an application globally, by formulating them as a set of Systems. We have implemented an experimental toolkit based on this approach, *Polyphony*, in order to demonstrate the use and benefits of this model.

7.4. Macro-dynamics

Participants: Stéphane Huot, Sylvain Malacria [correspondent], Nicole Pong.

One conspicuous feature of the current evolution of interactive devices is the spread of touch-sensitive surfaces. Typically, modern smartphones are equipped with such touch-sensitive surfaces that also support normal force-based input capabilities, which can for instance be used to control the range of a text selection by varying the force applied to the touchscreen (on e. g., iOS devices). However, this interaction mechanism is difficult to discover and many users simply ignore it exists. To overcome this problem, we introduced ForceSelect (see Figure 6 , left), a force-based text selection techniques that relies on a simple mode gauge (see Figure 6 , right) that does not require additional screen real-estate and help users to discover and master the use of force input in text selection tasks [22]. We conducted two studies that suggest that this mode gauge successfully provides enhanced discoverability of the force-based input and combines support for novices and experts, whereas it was never worse than the standard iOS technique and was also preferred by participants.

7.5. Interaction Machine

Several of our new results this year contributed to our global objective of building an Interaction Machine, especially at the micro-dynamics level. Our work on prediction algorithms and our hybrid hardware-software latency compensation method highlighted the need for accessing low-level input data and to have flexible input management to be able to reliably predict current finger position and compensate for latency. Our work on the characterization of the dimensions of touch interaction, especially angle of touch, highlighted the need for additional dimensions in input events that are not yet accessible in actual systems. All in all, this confirm our hypothesis that we have to redefine input management and input events propagation in order to better account for human factors in interactive systems, to extend the possibilities for designing more efficient and expressive interaction methods.

At the meso-dynamics level, our work on improving basic interaction methods in non-standard setups (e. g., VR, AR) highlighted the need for more open and flexible system architectures and tools that ease the design and prototyping of alternative interaction techniques based on mixed modalities. The new prototyping and programing tools that we proposed this year (StEM, ZenStates and Polyphony) are our first explorations toward such system-integrated frameworks dedicated to interaction.

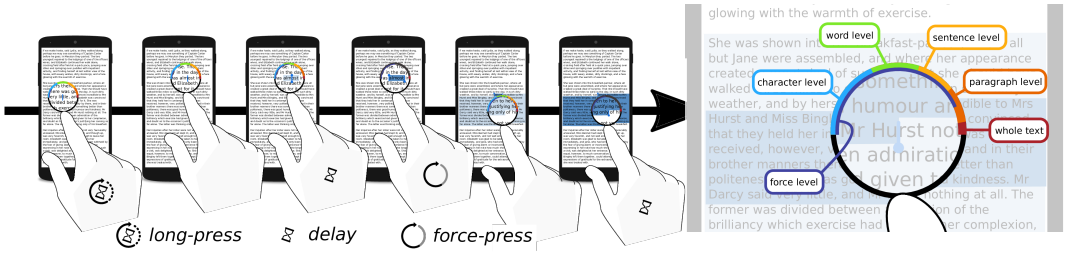

Figure 6. (left) Example of text selection using ForceSelect. The user performs a long-press that displays the callout magnifier. Keeping the force in the character level, the user adjusts its position by moving her finger. She then holds the force in the word level of the "mode gauge", locks the selection and enters the clutch mode. When force-pressing to the whole text level of the "mode gauge", she un-clutches the selection and updates it.; (right) Close-up of the "mode gauge". There are two types of text highlighting in the background: dark highlighting covers between both handles and light highlighting acts as a feedforward of which portion of text will be selected if the user released her finger (here the whole paragraph).

MAGNET Project-Team

7. New Results

7.1. On the Bernstein-Hoeffding Method

We consider extensions of Hoeffding's "exponential method" approach for obtaining upper estimates on the probability that a sum of independent and bounded random variables is significantly larger than its mean. We show that the exponential function in Hoeffding's approach can be replaced with any function which is non-negative, increasing and convex. As a result we generalize and improve upon Hoeffding's inequality. Our approach allows to obtain "missing factors" in Hoeffding's inequality. The later result is a rather weaker version of a theorem that is due to Michel Talagrand. Moreover, we characterize the class of functions with respect to which our method yields optimal concentration bounds. Finally, using ideas from the theory of Bernstein polynomials, we show that similar ideas apply under information on higher moments of the random variables ([4]).

7.2. IncGraph: Incremental graphlet counting for topology optimisation

Graphlets are small network patterns that can be counted in order to characterise the structure of a network (topology). As part of a topology optimisation process, one could use graphlet counts to iteratively modify a network and keep track of the graphlet counts, in order to achieve certain topological properties. Up until now, however, graphlets were not suited as a metric for performing topology optimisation; when millions of minor changes are made to the network structure it becomes computationally intractable to recalculate all the graphlet counts for each of the edge modifications. We propose IncGraph, a method for calculating the differences in graphlet counts with respect to the network in its previous state, which is much more efficient than calculating the graphlet occurrences from scratch at every edge modification made. In comparison to static counting approaches, our findings show IncGraph reduces the execution time by several orders of magnitude. The usefulness of this approach was demonstrated by developing a graphlet-based metric to optimise gene regulatory networks. IncGraph is able to quickly quantify the topological impact of small changes to a network, which opens novel research opportunities to study changes in topologies in evolving or online networks, or develop graphlet-based criteria for topology optimisation. IncGraph is freely available as an open-source R package on CRAN (incgraph). The development version is also available on GitHub (rcannood/incgraph) ([2]).

7.3. Graph sampling with applications to estimating the number of pattern embeddings and the parameters of a statistical relational model

Counting the number of times a pattern occurs in a database is a fundamental data mining problem. It is a subroutine in a diverse set of tasks ranging from pattern mining to supervised learning and probabilistic model learning. While a pattern and a database can take many forms, this paper focuses on the case where both the pattern and the database are graphs (networks). Unfortunately, in general, the problem of counting graph occurrences is #P-complete. In contrast to earlier work, which focused on exact counting for simple (i.e., very short) patterns, we present a sampling approach for estimating the statistics of larger graph pattern occurrences. We perform an empirical evaluation on synthetic and real-world data that validates the proposed algorithm, illustrates its practical behavior and provides insight into the trade-off between its accuracy of estimation and computational efficiency ([5]).

7.4. A machine learning based framework to identify and classify long terminal repeat retrotransposons

Transposable elements (TEs) are repetitive nucleotide sequences that make up a large portion of eukaryotic genomes. They can move and duplicate within a genome, increasing genome size and contributing to genetic diversity within and across species. Accurate identification and classification of TEs present in a genome is an important step towards understanding their effects on genes and their role in genome evolution. We introduce TE-LEARNER, a framework based on machine learning that automatically identifies TEs in a given genome and assigns a classification to them. We present an implementation of our framework towards LTR retrotransposons, a particular type of TEs characterized by having long terminal repeats (LTRs) at their boundaries. We evaluate the predictive performance of our framework on the well-annotated genomes of Drosophila melanogaster and Arabidopsis thaliana and we compare our results for three LTR retrotransposon superfamilies with the results of three widely used methods for TE identification or classification: REPEATMASKER, CENSOR and LTRDIGEST. In contrast to these methods, TE-LEARNER is the first to incorporate machine learning techniques, outperforming these methods in terms of predictive performance, while able to learn models and make predictions efficiently. Moreover, we show that our method was able to identify TEs that none of the above method could find, and we investigated TE-LEARNER's predictions which did not correspond to an official annotation. It turns out that many of these predictions are in fact strongly homologous to a known TE ([6]).

7.5. A Distributed Frank-Wolfe Framework for Learning Low-Rank Matrices with the Trace Norm

We consider the problem of learning a high-dimensional but low-rank matrix from a large-scale dataset distributed over several machines, where low-rankness is enforced by a convex trace norm constraint. We propose DFW-Trace, a distributed Frank-Wolfe algorithm which leverages the low-rank structure of its updates to achieve efficiency in time, memory and communication usage. The step at the heart of DFW-Trace is solved approximately using a distributed version of the power method. We provide a theoretical analysis of the convergence of DFW-Trace, showing that we can ensure sublinear convergence in expectation to an optimal solution with few power iterations per epoch. We implement DFW-Trace in the Apache Spark distributed programming framework and validate the usefulness of our approach on synthetic and real data, including the ImageNet dataset with high-dimensional features extracted from a deep neural network ([7]).

7.6. Personalized and Private Peer-to-Peer Machine Learning

The rise of connected personal devices together with privacy concerns call for machine learning algorithms capable of leveraging the data of a large number of agents to learn personalized models under strong privacy requirements. In this paper, we introduce an efficient algorithm to address the above problem in a fully decentralized (peer-to-peer) and asynchronous fashion, with provable convergence rate. We show how to make the algorithm differentially private to protect against the disclosure of information about the personal datasets, and formally analyze the trade-off between utility and privacy. Our experiments show that our approach dramatically outperforms previous work in the non-private case, and that under privacy constraints, we can significantly improve over models learned in isolation ([9]).

7.7. Hiding in the Crowd: A Massively Distributed Algorithm for Private Averaging with Malicious Adversaries

The amount of personal data collected in our everyday interactions with connected devices offers great opportunities for innovative services fueled by machine learning, as well as raises serious concerns for the privacy of individuals. In this paper, we propose a massively distributed protocol for a large set of users to privately compute averages over their joint data, which can then be used to learn predictive models. Our protocol can find a solution of arbitrary accuracy, does not rely on a third party and preserves the privacy of

users throughout the execution in both the honest-but-curious and malicious adversary models. Specifically, we prove that the information observed by the adversary (the set of malicious users) does not significantly reduce the uncertainty in its prediction of private values compared to its prior belief. The level of privacy protection depends on a quantity related to the Laplacian matrix of the network graph and generally improves with the size of the graph. Furthermore, we design a verification procedure which offers protection against malicious users joining the service with the goal of manipulating the outcome of the algorithm ([15]).

7.8. A Probabilistic Model for Joint Learning of Word Embeddings from Texts and Images

Several recent studies have shown the benefits of combining language and perception to infer word embeddings. These multimodal approaches either simply combine pre-trained textual and visual representations (e.g. features extracted from convolutional neural networks), or use the latter to bias the learning of textual word embeddings. In this work, we propose a novel probabilistic model to formalize how linguistic and perceptual inputs can work in concert to explain the observed word-context pairs in a text corpus. Our approach learns textual and visual representations jointly: latent visual factors couple together a skip-gram model for co-occurrence in linguistic data and a generative latent variable model for visual data. Extensive experimental studies validate the proposed model. Concretely, on the tasks of assessing pairwise word similarity and image/caption retrieval, our approach attains equally competitive or stronger results when compared to other state-of-the-art multimodal models ([8]).

7.9. A Framework for Understanding the Role of Morphology in Universal Dependency Parsing

We present a simple framework for characterizing morphological complexity and how it encodes syntactic information. In particular, we propose a new measure of morpho-syntactic complexity in terms of governor-dependent preferential attachment that explains parsing performance. Through experiments on dependency parsing with data from Universal Dependencies (UD), we show that representations derived from morphological attributes deliver important parsing performance improvements over standard word form embeddings when trained on the same datasets. We also show that the new morpho-syntactic complexity measure is predictive of the gains provided by using morphological attributes over plain forms on parsing scores, making it a tool to distinguish languages using morphology as a syntactic marker from others ([11]).

7.10. Online Reciprocal Recommendation with Theoretical Performance Guarantees

A reciprocal recommendation problem is one where the goal of learning is not just to predict a user's preference towards a passive item (e.g., a book), but to recommend the targeted user on one side another user from the other side such that a mutual interest between the two exists. The problem thus is sharply different from the more traditional items-to-users recommendation, since a good match requires meeting the preferences at both sides. We initiate a rigorous theoretical investigation of the reciprocal recommendation task in a specific framework of sequential learning. We point out general limitations, formulate reasonable assumptions enabling effective learning and, under these assumptions, we design and analyze a computationally efficient algorithm that uncovers mutual likes at a pace comparable to that achieved by a clairvoyant algorithm knowing all user preferences in advance. Finally, we validate our algorithm against synthetic and real-world datasets, showing improved empirical performance over simple baselines ([13]).

7.11. On Similarity Prediction and Pairwise Clustering

We consider the problem of clustering a finite set of items from pairwise similarity information. Unlike what is done in the literature on this subject, we do so in a passive learning setting, and with no specific constraints on the cluster shapes other than their size. We investigate the problem in different settings: i. an

online setting, where we provide a tight characterization of the prediction complexity in the mistake bound model, and ii. a standard stochastic batch setting, where we give tight upper and lower bounds on the achievable generalization error. Prediction performance is measured both in terms of the ability to recover the similarity function encoding the hidden clustering and in terms of how well we classify each item within the set. The proposed algorithms are time efficient ([12]).

7.12. A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization

The performance of many machine learning techniques depends on the choice of an appropriate similarity or distance measure on the input space. Similarity learning (or metric learning) aims at building such a measure from training data so that observations with the same (resp. different) label are as close (resp. far) as possible. In this paper, similarity learning is investigated from the perspective of pairwise bipartite ranking, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point, based on the similarity scores. A natural performance criterion in this setting is pointwise ROC optimization: maximize the true positive rate under a fixed false positive rate. We study this novel perspective on similarity learning through a rigorous probabilistic framework. The empirical version of the problem gives rise to a constrained optimization formulation involving U-statistics, for which we derive universal learning rates as well as faster rates under a noise assumption on the data distribution. We also address the large-scale setting by analyzing the effect of sampling-based approximations. Our theoretical results are supported by illustrative numerical experiments ([14]).

7.13. Escaping the Curse of Dimensionality in Similarity Learning: Efficient Frank-Wolfe Algorithm and Generalization Bounds

Similarity and metric learning provides a principled approach to construct a task-specific similarity from weakly supervised data. However, these methods are subject to the curse of dimensionality: as the number of features grows large, poor generalization is to be expected and training becomes intractable due to high computational and memory costs. In this paper, we propose a similarity learning method that can efficiently deal with high-dimensional sparse data. This is achieved through a parameterization of similarity functions by convex combinations of sparse rank-one matrices, together with the use of a greedy approximate Frank-Wolfe algorithm which provides an efficient way to control the number of active features. We show that the convergence rate of the algorithm, as well as its time and memory complexity, are independent of the data dimension. We further provide a theoretical justification of our modeling choices through an analysis of the generalization error, which depends logarithmically on the sparsity of the solution rather than on the number of features. Our experiments on datasets with up to one million features demonstrate the ability of our approach to generalize well despite the high dimensionality as well as its superiority compared to several competing methods ([16]).

7.14. Nonstochastic Bandits with Composite Anonymous Feedback

We investigate a nonstochastic bandit setting in which the loss of an action is not immediately charged to the player, but rather spread over at most d consecutive steps in an adversarial way. This implies that the instantaneous loss observed by the player at the end of each round is a sum of as many as d loss components of previously played actions. Hence, unlike the standard bandit setting with delayed feedback, here the player cannot observe the individual delayed losses, but only their sum. Our main contribution is a general reduction transforming a standard bandit algorithm into one that can operate in this harder setting. We also show how the regret of the transformed algorithm can be bounded in terms of the regret of the original algorithm. Our reduction cannot be improved in general: we prove a lower bound on the regret of any bandit algorithm in this setting that matches (up to log factors) the upper bound obtained via our reduction. Finally, we show how our reduction can be extended to more complex bandit settings, such as combinatorial linear bandits and online bandit convex optimization ([10]).

MEPHYSTO-POST Team

4. New Results

4.1. Exponential time-decay for discrete Fokker–Planck equations

G. Dujardin and his coauthors proposed and studied in [22] several discrete versions of homogeneous and inhomogeneous one-dimensional Fokker-Planck equations. They proved in particular, for these discretizations of velocity and space, the exponential convergence to the equilibrium of the solutions, for time-continuous equations as well as for time-discrete equations. Their method uses new types of discrete Poincaré inequalities for a "two-direction" discretization of the derivative in velocity. For the inhomogeneous problem, they adapted hypocoercive methods to the discrete level.

4.2. Energy preserving methods for nonlinear Schrödinger equations

G. Dujardin and his coauthors have revisited and extended relaxation methods for nonlinear Schrödinger equations (NLS). The classical relaxation method for NLS is an energy preserving method and a mass preserving method. Moreover, it is only linearly implicit. A first proof of the second order accuracy was achieved in [14]. Moreover, the method was extended to enable to treat noncubic nonlinearities, nonlocal nonlinearities, as well as rotation terms. The resulting methods are still energy preserving and mass preserving. Moreover, they are shown to have second order accuracy numerically. These new methods are compared with fully implicit, mass and energy preserving methods of Crank and Nicolson.

4.3. Diffusive and superdiffusive behavior in one-dimensional chains of oscillators

In order to understand abnormally diffusive phenomena which are physically observed in nanotube technologies, one mathematical approach consists in starting from deterministic system of Newtonian particles, and then perturb this system with a stochastic component which provides enough ergodicity to the dynamics. It is already well known that these stochastic chains model correctly the behavior of the conductivity [24]. In [1], [2] (published in Communications in Mathematical Physics) M. Simon with her coauthors C. Bernardin, P. Gonçalves, M. Jara, T. Komorowski, S. Olla and M. Sasada have observed both behaviors, normal and anomalous diffusion, in the context of low dimensional asymmetric systems. They manage to describe the microscopic phenomena at play which are responsible for each one of these phenomena, and they go beyond the predictions that have recently been done in [31], [32]. Moreover, in [2], the authors manage to treat rigorously, for the first time, the case of an anharmonic potential: more precisely, they consider a small quartic anharmonicity and show that the result obtained in the harmonic (linear) case persists up to some small critical value of the nonlinear perturbation.

4.4. Microscopic description of moving interfaces

A large variety of models has been introduced to describe the evolution of a multiphase medium, *e.g.* the joint evolution of liquid and solid phases. These complex physical phenomena often feature absorbing phase transitions. For instance, the porous medium equation (PME)

$$\partial_t \rho = \mathrm{div}(\rho^{m-1}\, \nabla \rho), \tag{1}$$

where $m > 1$ is a constant and div and ∇ are the divergence and gradient operators in $\mathbb{R}^d$, describes the evolution of the density $\rho : \mathbb{R}^d \times \mathbb{R}_+ \to [0, 1]$ of an ideal gas flowing in a homogeneous medium. It is known that, starting from an initial density ρ_0 with compact support, the solution $\rho(x, t)$ is nonnegative and has compact support in the space variable for each positive t. Thus there are interfaces separating the regions where ρ is positive from those where it is zero.

In one submitted paper in collaboration with O. Blondel, C. Cancès, and M. Sasada, we have derived the PME (1) from a degenerate and conservative dynamics in [15], for any integer $m > 1$. More precisely we improved the results previously obtained in [26], since we allow the solutions to feature moving interfaces, namely the initial condition may vanish. This moving boundary was not well apprehended at the microscopic level. Its rigorous definition is indeed very delicate, and its behavior (such that its speed, or fluctuation), as well as the relationship between the microscopic and macroscopic boundaries, are challenging questions that we aim to tackle in a near future.

When $m < 1$, equation (1) is called fast diffusion equation. In a recent collaborative work (submitted) with O. Blondel, C. Erignoux and M. Sasada [16], we derive such a fast diffusion equation in dimension one from an interacting particle system belonging to the class of conserved lattice gases with active-absorbing phase transition [30]. The microscopic dynamics is very constrained: in a few words, a particle can jump to the right (resp. left) empty neighboring site if and only if it has a particle to its left (resp. right) neighboring site. This model is really complex: the state space is divided into transient states, absorbing states and ergodic states. Depending on the initial number of particles, the transient good configurations will lead to the ergodic component and the transient bad configurations will be absorbed to an inactive state. Because of the jump constraint, there are two distinct regimes for the macroscopic behavior. Either the macroscopic density is larger than $\frac{1}{2}$, in which case the system behaves diffusively, or the density is lower than $\frac{1}{2}$, in which case the system freezes rapidly.

The interfaces between these two phases propagate as particles from the supercritical phase ($\rho > \frac{1}{2}$) diffuse towards the subcritical phase ($\rho < \frac{1}{2}$). We expect that the macroscopic density profile evolves under the diffusive scaling according to the Stefan problem

$$\partial_t \rho = \Delta\left(G(\rho)\right) \qquad \text{where } G(\rho) = \tfrac{2\rho-1}{\rho}\, \mathbf{1}_{\rho>\frac{1}{2}}. \tag{2}$$

The microscopic derivation of such Stefan problems is a well known difficult problem, only partially solved [27], [29]. In [16] we treat the liquid part of the problem (*i.e.* when the initial profiles ρ_0 are uniformly larger than the critical density $\frac{1}{2}$) and we provide a refined estimation of the time needed by the system to enter into the ergodic state. Then, we show that the macroscopic density profile evolves under the diffusive time scaling according to (1) with $m = -1$. The extension to more general initial profiles is our next goal.

4.5. Stability analysis of a Vlasov-Wave system

S. De Bièvre and his co-authors introduced and studied a kinetic equation of the Vlasov-Wave type, which arises in the description of the behavior of a large number of particles interacting weakly with an environment, composed of an infinite collection of local vibrational degrees of freedom, modeled by wave equations. They use variational techniques to establish the existence of large families of stationary states for this system, and analyze their stability [8].

4.6. Orbital stability in the presence of symmetries

With S. Rota Nodari, S. De Bièvre considered the orbital stability of relative equilibria of Hamiltonian dynamical systems on Banach spaces, in the presence of a multi-dimensional invariance group for the dynamics [9]. They proved a persistence result for such relative equilibria, presented a generalization of the Vakhitov-Kolokolov slope condition to this higher dimensional setting, and showed how it allows to prove the local coercivity of the Lyapunov function, which in turn implies orbital stability. The method was applied to study the orbital stability of relative equilibria of nonlinear Schrödinger and Manakov equations. It extends and clarifies the approach of Grillakis-Shatah-Strauss.

4.7. Measuring nonclassicality of bosonic field quantum state

S. De Bièvre and his collaborators introduced a new distance-based measure for the nonclassicality of the states of a bosonic field, which outperforms the existing such measures in several ways [17]. They defined for that purpose the operator ordering sensitivity of the state which evaluates the sensitivity to operator ordering of the Renyi entropy of its quasi-probabilities and which measures the oscillations in its Wigner function. Through a sharp control on the operator ordering sensitivity of classical states they obtained a precise geometric image of their location in the density matrix space allowing them to introduce a distance-based measure of nonclassicality. They analyze the link between this nonclassicality measure and a recently introduced quantum macroscopicity measure, showing how the two notions are distinct.

4.8. The Cauchy problem for the Landau–Lifshitz–Gilbert equation in BMO and self-similar solutions

A. de Laire and S. Gutierrez established in [19] a global well-posedness result for the Landau–Lifshitz equation with Gilbert damping, provided that the BMO semi-norm of the initial data is small. As a consequence, they deduced the existence of self-similar solutions in any dimension. Moreover, in the one-dimensional case, they characterized the self-similar solutions when the initial data is given by some ($\S^2$-valued) step function and established their stability. They also showed the existence of multiple solutions if the damping is strong enough.

4.9. The Sine–Gordon regime of the Landau–Lifshitz equation with a strong easy-plane anisotropy

It is well-known that the dynamics of biaxial ferromagnets with a strong easy-plane anisotropy is essentially governed by the Sine-Gordon equation. A. de Laire and P. Gravejat provided in [10] a rigorous justification to this observation. More precisely, they showed the convergence of the solutions to the Landau-Lifshitz equation for biaxial ferromagnets towards the solutions to the Sine-Gordon equation in the regime of a strong easy-plane anisotropy. This result holds for solutions to the Landau–Lifshitz equation in high order Sobolev spaces. They also provided an alternative proof for local well-posedness in this setting by introducing high order energy quantities with better symmetrization properties. Then they derived the convergence from the consistency of the Landau–Lifshitz equation with the Sine-Gordon equation by using well-tailored energy estimates. As a by-product, they also obtained a further derivation of the free wave regime of the Landau–Lifshitz equation.

4.10. Mutual information of wireless channels and block-Jacobi ergodic operators

In telecommunication models the quality of the transferred data is assessed through the entropy of the channel, a theoretical quantity that is usually not computable in practice. W. Hachem, A. Hardy and S. Shamai prove in [23] that one can relate this quantity for a large class of models involving several antennas (MIMO) to the equilibrium measure of a matrix valued Markov chain associated with the model, and so does its asymptotic behavior when the signal-noise-ratio parameter becomes large. By means of ergodicity results, this yields estimates for these quantities that are implementable faster than the naive estimators.

4.11. DLR equations and rigidity for the Sine-beta process

The Sine-beta process is a universal object appearing in the study of large Hermitian random matrices and statistical systems in a logarithmic interaction, such as low dimensional Coulomb gases. However, the only description available yet relied on a rather complicated and non-physical system of coupled stochastic differential equations. In [21], D. Dereudre, A. Hardy, T. Leblé and M. Maïda obtain a statistical physic interpretation of the Sine-beta process as probability measure on infinite configurations of points described by means of the DLR formalism. This allows to obtain more information on the Sine-beta process: for instance, it is rigid, it is tolerant, and the number of particles in a compact box has gaussian fluctuations as the box becomes large.

4.12. Time-frequency transforms of white noises and Gaussian analytic functions

In signal processing, an important challenge is to be able to separate signals from ambient noises. In time-frequency analysis, this problem reduces to identify what is the spectrogram of a white noise to derive statistical tests in order to decide if some partial signal is noise or not. P. Fandrin recently put forward that the understanding of the zeros of the spectrograms would be already an important step by analyticity of the spectrograms. R. Bardenet and A. Hardy observed in [13] that there is a canonical way to identify the zeros of the usual white noise transforms associated to classical spectrograms and zeroes of Gaussian analytic functions associated with classical orthogonal polynomials in the background. In particular the zeros satisfy some invariance properties leading to computable correlation functions. In specific cases, one can identify some transforms whose zeros form a determinantal point process, in which case all the statistics of interests can be computed explicitly and this allows an exact numerical treatment.

4.13. Energy of the Coulomb gas on the sphere at low temperature

In relation to the 7th Smale problem, which is about finding polynomial time algorithm to produce well spread configuration of points on the sphere in a quantified manner, C. Beltran and A. Hardy proved in [4] that the Coulomb gas on the sphere at a temperature proportional to the inverse number of points in a configuration reaches the numerical precision required by this problem. We however did not discuss yet the algorithmic procedure, which is currently in investigation by A. Hardy and M. Simon.

4.14. Polynomial ensembles and recurrence coefficients

Determinantal point processes can be of important use in applications as soon as one is interested in producing configurations of well spread points on an arbitrary space. A class of determinantal point processes on the real line that has been extensively studied recently are the so-called polynomial ensembles. A. Hardy gathered in [11] several results concerning these models in relation to the recurrent coefficients associated with the orthogonal polynomials hidden in the background.

4.15. Concentration for Coulomb gases and Coulomb transport inequalities

The convergence of the Coulomb gas, which is a statistical gas of charged particles in an electrostatic interaction, towards its limiting distribution as the number of particles goes to infinity is a result which is part of the folklore of potential theory. The speed at which this convergence arise, which can be assessed through concentration of measure estimates in, say, the Wasserstein-Kantorovich metric, are however new results obtained by D. Chafaï, A. Hardy and M. Maïda in [7]. One of the main ingredient was to develop transport inequalities associated with the Coulomb interaction.

MODAL Project-Team

7. New Results

7.1. Axis 1: Data Units Selection in Statistics

Participant: Christophe Biernacki.

Usually, the data unit definition is fixed by the practitioner but it can happen that he/her hesitates between several data unit options. In this context, it is highlighted that it is possible to embed data unit selection into a classical model selection principle. The problem is introduced in a regression context before to focus on the model-based clustering and co-clustering context, for data of different kinds (continuous, count, categorical). This work is now published in an international journal [12].

An extension of this work has been also presented to an international workshop.The idea is to use the data units principle as a way for (co-)clustering model enlargement.

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.2. Axis 1: Model-Based Co-clustering for Ordinal Data

Participant: Christophe Biernacki.

A model-based co-clustering algorithm for ordinal data is presented. This algorithm relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Model inference relies on a Stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of co-clusters (or blocks). The main advantage of this ordinal dedicated co-clustering model is its parsimony, the interpretability of the co-cluster parameters (mode, precision) and the possibility to take into account missing data. Numerical experiments on simulated data show the efficiency of the inference strategy, and real data analyses illustrate the interest of the proposed procedure. The resulting work is now published in the international journal [18]. This is joint work Julien Jacques from University of Lyon 2.

7.3. Axis 1: Model-Based Co-clustering for Ordinal Data of different dimensions

Participant: Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological position, it is also useful to observe how questions may be grouped. This is why a clustering should also be performed on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work is now under revision in an international journal [54] and has been presented to an international conference [38]. The related R package ordinalClust has been also written and has led to a specific preprint [57].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2, and Florence Cousson-Gélie from University Paul Valéry Montpellier 3.

7.4. Axis 1: Model-based co-clustering for mixed type data

Participant: Christophe Biernacki.

Over decades, a lot of studies have shown the importance of clustering to emphasize groups of observations. More recently, due to the emergence of high-dimensional datasets with a huge number of features, co-clustering techniques have emerged and proposed several methods for simultaneously producing groups of observations and features. By synthesizing the dataset in blocks (the crossing of a row-cluster and a column-cluster), this technique can sometimes summarize better the data and its inherent structure. The Latent Block Model (LBM) is a well-known method for performing a co-clustering. However, recently, contexts with features of different types (here called mixed type datasets) are becoming more common. Unfortunately, the LBM is not directly applicable on this kind of dataset. The present work extends the usual LBM to the so-called Multiple Latent Block Model (MLBM) which is able to handle mixed type datasets. The inference is done through a Stochastic EM-algorithm embedding a Gibbs sampler and model selection criterion is defined to choose the number of row and column clusters. This method was successfully used on simulated and real datasets. This work is available as a preprint [55] which has been submitted to an international journal. It has also led to the R package mixedClust which has been presented to an international workshop [56] and has led to a specific preprint [56].

An adaptation of this general principle to the specific case of mixing textual and continuous data has been also proposed and presented to a national conference [26], with an international audience.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.5. Axis 1: Model-Based Co-clustering with Co-variables

Participant: Serge Iovleff.

This work has been motivated by an epidemiological and genetic survey of malaria disease in Senegal. Data were collected between 1990 and 2008. It is based on a latent block model taking into account the problem of grouping variables and clustering individuals by integrating information given by a set of co-variables. Numerical experiments on simulated data sets and an application on real genetic data highlight the interest of this approach. BEM algorithm is deduced and implemented in R package *simerge* and has led to a specific preprint [24].

7.6. Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-Clustering for High Dimensional Data

Participant: Christophe Biernacki.

A co-clustering model for continuous data that relaxes the identically distributed assumption within blocks of traditional co-clustering is presented. The proposed model, although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering. A stochastic EM algorithm along with a Gibbs sampler is used for parameter estimation and an ICL criterion is used for model selection. Simulated and real datasets are used for illustr1ation and comparison with traditional co-clustering. This work has led to a preprint

This is a joint work with Michael Gallaugher (PhD student) and Paul Mcnicholas, both from McMaster University (Canada). Michael Gallaughter visited the Modal for three months in 2018.

7.7. Axis 1: Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work has been submitted to an international journal [12] and has also been presented to an international conference [41].

This is a joint work with Matthieu Marbac from ENSAI.

7.8. Axis 1: A targeted multi-partitions clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

Clustering is generally not a purpose by itself, because its results are mainly tools used by the statistician for another analysis. Indeed, in many applications, clusters are assessed from a set of observed variables, then these clusters are used to predict other variables which are used or not in clustering. Because the final objective of prediction is not considered during cluster analysis, there is no reason to obtain relevant clusters for the variables to predict. We present a unified approach which simultaneously performs cluster analysis and prediction. This method considers that the variables to clusters arise from a product of finite mixture models which provides multiple partition. Moreover, the variables to predict are considered to be independent of the variables to cluster given the partition. The predictions are achieved by a generalized linear model. Model selection is conducted by optimizing the BIC. This optimization is achieved with a modified version of the EM algorithm which performs model selection and maximum likelihood inference simultaneously. An early version of this work has been presented to an international conference [37].

It is a joint work with Matthieu Marbac from ENSAI and with Mohamed Sedki from Université Paris-Sud.

7.9. Axis 1: Co-clustering: A versatile way to perform clustering in high dimension

Participant: Christophe Biernacki.

Standard model-based clustering is known to be very efficient for low-dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that co-clustering, an unsupervised mixture model learning method to define simultaneously groups of rows (individuals) and groups of columns (variables) on a data matrix, is of particular interest to perform HD clustering of individuals even if it is not its primary mission. Indeed, column clustering is recast as a strategy to control the variance of the estimation, the model dimension being driven by the number of groups of variables instead of the number of variables itself. However, the statistical counterpart of this important variance reduction brings naturally some important model bias. The purpose is to access (first in an empirical manner) the trade-off bias-variance of the co-clustering strategy in scenarios involving HD fundaments (correlated variables, irrelevant variables). We show the ability of co-clustering to outperform simple mixture row-clustering, even if co-clustering clearly corresponds to a misspecified model situation, revealing a promising manner to efficiently address (very) HD clustering. An early version of this work has been presented to an international conference [36].

It is a joint work with Christine Keribin from Université Paris-Sud.

7.10. Axis 1: Dealing with missing data in model-based clustering through a MNAR model

Participants: Christophe Biernacki, Fabien Laporte.

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case, some information is lost; in the second case, the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data. An early version of this work has been presented to an international conference [33].

It is a joint work with Gilles Celeux from Inria Saclay and Julie Josse from Ecole Polytechnique.

7.11. Axis 1: Self Organizing Coclustering for textual data synthesis

Participant: Christophe Biernacki.

Recently, different studies have demonstrated the interest of co-clustering, which simultaneously produces clusters of lines and columns. The present work introduces a novel co-clustering model for parsimoniously summarizing textual data in documents × terms format. Besides highlighting homogeneous coclusters - as other existing algorithms do - we also distinguish noisy coclusters from significant ones, which is particularly useful for sparse documents × term matrices. Furthermore, our model proposes a structure among the significant coclusters and thus obtains a better interpretability to the user. By forcing a structure through row-clusters and column-clusters, this approach is competitive in terms of documents clustering, and offers user-friendly results. The algorithm derived for the proposed method is a Stochastic EM algorithm embedding a Gibbs sampling step and the Poisson distribution. A preprint is currently in progress.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.12. Axis 1: Linking canonical and spectral clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

It is a recent work aiming at defining a mathematical bridge between classical model-based clustering and classical spectral clustering. Interest of such a prospect is to be able to compare both methods through the rigorous scheme of model selection paradigm. It is an ongoing work.

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.13. Axis 1: Multiple partition clustering

Participant: Vincent Vandewalle.

In the framework of model-based clustering, a model allowing several latent class variables have been proposed. This model assumes that the distribution of the observed data can be factorized into several independent blocks of variables. Each block is assumed to follow a latent class model (i.e., mixture with conditional independence assumption). The proposed model includes variable selection, as a special case, and is able to cope with the mixed-data setting. The simplicity of the model allows to estimate the repartition of the variables into blocks and the mixture parameters simultaneously, thus avoiding running EM algorithms for each possible repartition of variables into blocks. For the proposed method, a model is defined by the number of blocks, the number of clusters inside each block and the repartition of variables into blocks. Model selection can be done with two information criteria, the BIC and the MICL, for which an efficient optimization is proposed. The proposed method gives a rich interpretation of the data set at hand (i.e., analysis of the repartition of the variables into blocks and analysis of the clusters produced by each block of variables). This work as been presented in several international conferences and is now published [20].

It is a joint work with Matthieu Marbac from ENSAI.

7.14. Axis 2: Change-point detection by means of reproducing kernels

Participant: Alain Celisse.

Classical offline change-point detection approaches are limited to detecting changes arising in the mean and/or variance of the distribution along the time. Detecting changes in other moments of the distribution is possible, but at the price of stronger (unrealistic) distributional assumptions which are likely to be violated.

Reproducing kernels are a means to detect changes arising in any moments of the distribution along the time, which are not limited to the mean or the variance. One of the main contributions of this work is to provide a theoretically grounded model selection strategy allowing us to detection multiple changes. From additional extensive simulation experiments, it clearly arises that the so-called KCP approach outperforms numerous state-of-the art change-points detection procedures such as E-divisive, PELT, ...

7.15. Axis 2: New efficient algorithms for multiple change-point detection with kernels

Participants: Alain Celisse, Guillemette Marot.

Several statistical approaches based on reproducing kernels have been proposed to detect abrupt changes arising in the full distribution of the observations and not only in the mean or variance. Some of these approaches enjoy good statistical properties (oracle inequality, ...). Nonetheless, they have a high computational cost both in terms of time and memory. This makes their application difficult even for small and medium sample sizes ($n < 10^4$). This computational issue is addressed by first describing a new efficient and exact algorithm for kernel multiple change-point detection with an improved worst-case complexity that is quadratic in time and linear in space. It allows dealing with medium size signals (up to $n \approx 10^5$). Second, a faster but approximation algorithm is described. It is based on a low-rank approximation to the Gram matrix. It is linear in time and space. This approximation algorithm can be applied to large-scale signals ($n \geq 10^6$). These exact and approximation algorithms have been implemented in `R` and `C` for various kernels. The computational and statistical performances of these new algorithms have been assessed through empirical experiments. The runtime of the new algorithms is observed to be faster than that of other considered procedures. Finally, simulations confirmed the higher statistical accuracy of kernel-based approaches to detect changes that are not only in the mean. These simulations also illustrate the flexibility of kernel-based approaches to analyze complex biological profiles made of DNA copy number and allele B frequencies. An R package implementing the approach will be made available on github.

7.16. Axis 2: Multi-Layer Group-Lasso

Participants: Alain Celisse, Guillemette Marot.

Multi-Layer Group-Lasso (MLGL) is a new procedure of variable selection in the context of redundancy between explanatory variables, which holds true with high-dimensional data. A sparsity assumption is made that is, only a few variables are assumed to be relevant for predicting the response variable. In this context, the performance of classical Lasso-based approaches strongly deteriorate as the redundancy strengthens. The proposed approach combines variable aggregation and selection in order to improve interpretability and performance. First, a hierarchical clustering procedure provides at each level a partition of the variables into groups. Then, the set of groups of variables from the different levels of the hierarchy is given as input to group-Lasso, with weights adapted to the structure of the hierarchy. At this step, group-Lasso outputs sets of candidate groups of variables for each value of regularization parameter. The versatility offered by MLGL to choose groups at different levels of the hierarchy a priori induces a high computational complexity. MLGL however exploits the structure of the hierarchy and the weights used in group-lasso to greatly reduce the final time cost. The final choice of the regularization parameter – and therefore the final choice of groups – is made by a multiple hierarchical testing procedures. A paper associated to the R package MLGL has been submitted [45].

7.17. Axis 2: Pseudo-Bayesian Learning with Kernel Fourier Transform as Prior

Participants: Pascal Germain, Gael Letarte.

We revisit the kernel random Fourier features (RFF) method through the lens of the PAC-Bayesian theory. While the primary goal of RFF is to approximate a kernel, we look at the Fourier transform as a prior distribution over trigonometric hypotheses. It naturally suggests learning a posterior on these hypotheses. We derive generalization bounds that are optimized by learning a pseudo-posterior obtained from a closed-form expression, and corresponding learning algorithms This work has been accepted for publication at AISTATS 2019 conference [51].

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne.

7.18. Axis 2: Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles

Participant: Benjamin Guedj.

We examine a network of learners which address the same classification task but must learn from different data sets. The learners can share a limited portion of their data sets so as to preserve the network load. We introduce DELCO (standing for Decentralized Ensemble Learning with COpulas), a new approach in which the shared data and the trained models are sent to a central machine that allows to build an ensemble of classifiers. The proposed method aggregates the base classifiers using a probabilistic model relying on Gaussian copulas. Experiments on logistic regressor ensembles demonstrate competing accuracy and increased robustness as compared to gold standard approaches. A companion python implementation can be downloaded at https://github.com/john-klein/DELCO.

Joint work with John Klein, Olivier Colot, Mahmoud Albardan (all from CRIStAL lab, UMR 9189, Univ. Lille. Preprint submitted: [50].

7.19. Axis 2: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Participants: Benjamin Guedj, Le Li.

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. Principal curves act as a nonlinear generalization of PCA and the present paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret bound and performance on a toy example and seismic data.

Preprint submitted: [47].

7.20. Axis 2: A Quasi-Bayesian Perspective to Online Clustering

Participants: Benjamin Guedj, Le Li.

When faced with high frequency streams of data, clustering raises theoretical and algorithmic pitfalls. We introduce a new and adaptive online clustering algorithm relying on a quasi-Bayesian approach, with a dynamic (i.e., time-dependent) estimation of the (unknown and changing) number of clusters. We prove that our approach is supported by minimax regret bounds. We also provide an RJMCMC-flavored implementation (called PACBO, see https://cran.r-project.org/web/packages/PACBO/index.html) for which we give a convergence guarantee. Finally, numerical experiments illustrate the potential of our procedure.

Joint work with Sébastien Loustau (LumenAI). Paper published in Electronic Journal of Statistics: https://projecteuclid.org/euclid.ejs/1537430425, [19].

7.21. Axis 2: Pycobra: A Python Toolbox for Ensemble Learning and Visualisation

Participants: Benjamin Guedj, Bhargav Srinivasa Desikan.

We introduce pycobra, a Python library devoted to ensemble learning (regression and classification) and visualisation. Its main assets are the implementation of several ensemble learning algorithms, a flexible and generic interface to compare and blend any existing machine learning algorithm available in Python libraries (as long as a predict method is given), and visualisation tools such as Voronoi tessellations. pycobra is fully scikit-learn compatible and is released under the MIT open-source license. pycobra can be downloaded from the Python Package Index (PyPi) and Machine Learning Open Source Software (MLOSS). The current version (along with Jupyter notebooks, extensive documentation, and continuous integration tests) is available at https://github.com/bhargavvader/pycobra and official documentation website is https://modal.lille.inria.fr/pycobra.

Paper published in Journal of Machine Learning Research: http://jmlr.org/papers/v18/17-228.html, [17]. Software submitted to the `scikit-learn-contrib` repository (under review).

7.22. Axis 2: Simpler PAC-Bayesian bounds for hostile data

Participant: Benjamin Guedj.

PAC-Bayesian learning bounds are of the utmost interest to the learning community. Their role is to connect the generalization ability of an aggregation distribution ρ to its empirical risk and to its Kullback-Leibler divergence with respect to some prior distribution π. Unfortunately, most of the available bounds typically rely on heavy assumptions such as boundedness and independence of the observations. This paper aims at relaxing these constraints and provides PAC-Bayesian learning bounds that hold for dependent, heavy-tailed observations (hereafter referred to as hostile data). In these bounds the Kullack-Leibler divergence is replaced with a general version of Csiszár's f-divergence. We prove a general PAC-Bayesian bound, and show how to use it in various hostile settings.

Joint work with Pierre Alquier (ENSAE ParisTech). Paper published in Machine Learning: [11].

7.23. Axis 2: PAC-Bayesian high dimensional bipartite ranking

Participant: Benjamin Guedj.

This paper is devoted to the bipartite ranking problem, a classical statistical learning task, in a high dimensional setting. We propose a scoring and ranking strategy based on the PAC-Bayesian approach. We consider nonlinear additive scoring functions, and we derive non-asymptotic risk bounds under a sparsity assumption. In particular, oracle inequalities in probability holding under a margin condition assess the performance of our procedure, and prove its minimax optimality. An MCMC-flavored algorithm is proposed to implement our method, along with its behavior on synthetic and real-life datasets.

Joint work with Sylvain Robbiano. Paper published in Journal of Statistical Planning and Inference: [16].

7.24. Axis 2: Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters

Participant: Pascal Germain.

We propose a boosting based multiview learning algorithm which iteratively learns i) weights over view-specific voters capturing view-specific information; and ii) weights over views by optimizing a PAC-Bayes multiview C-Bound that takes into account the accuracy of view-specific classifiers and the diversity between the views. We derive a generalization bound for this strategy following the PAC-Bayes theory which is a suitable tool to deal with models expressed as weighted combination over a set of voters. This work has been submitted to an international journal and is available as a preprint [44].

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne and with Massih-Reza Amini of Université de Grenoble, and with Anil Goyal affiliated to both institutions.

7.25. Axis 3: Clustering spatial functional data

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle.

We propose two approaches for clustering spatial functional data. The first one is the model-based clustering that uses the concept of density for functional random variables. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean. These two approaches take into account the spatial features of the data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data. This work will appear in the "Geostatistical Functional Data Analysis: Theory and Methods". Wiley, 2018. Editors : Jorge Mateu, Ramon Giraldo [39].

7.26. Axis 3: Categorical functional data analysis

Participants: Cristian Preda, Vincent Vandewalle.

We develop and implement techniques for analysis of categorical functional data. Visualization, clustering and regression methods with categorical functional predictor are proposed.

7.27. Axis 4: Real-time Audio Sources Classification

Participants: Christophe Biernacki, Maxime Baelde.

This work addresses the recurring challenge of real-time monophonic and polyphonic audio source classification. The whole power spectrum is directly involved in the proposed process, avoiding complex and hazardous traditional feature extraction. It is also a natural candidate for polyphonic events thanks to its additive property in such cases. The classification task is performed through a nonparametric kernel-based generative modeling of the power spectrum. Advantage of this model is twofold: it is almost hypothesis free and it allows to straightforwardly obtain the maximum a posteriori classification rule of online signals. Moreover it makes use of the monophonic dataset to build the polyphonic one. Then, to reach the real-time target, the complexity of the method can be tuned by using a standard hierarchical clustering preprocessing of sound models, revealing a particularly efficient computation time and classification accuracy trade-off. The proposed method reveals encouraging results both in monophonic and polyphonic classification tasks on benchmark and owned datasets, even in real-time situations. This method also has several advantages compared to the state-of-the-art methods include a reduced training time, no hyperparameters tuning, the ability to control the computation - accuracy trade-off and no training on already mixed sounds for polyphonic classification. This work is now under revision to an international journal [40].

It is a joint work with Raphaël Greff, from the A-Volute company.

7.28. Axis 4: Matching of descriptors evolving over time

Participants: Christophe Biernacki, Anne-Lise Bedenel.

In the web domain, and in particular for insurance comparison, data constantly evolve, implying that it is difficult to directly exploit them. For example, to do a classification, performing standard learning processes require data descriptors equal for both learning and test samples. Indeed, for answering web surfer expectation, online forms whence data come from are regularly modified. So, features and data descriptors are also regularly modified. In this work, it is introduced a process to estimate and understand connections between transformed data descriptors. This estimated matching between descriptors will be a preliminary step before applying later classical learning methods. This work has been presented to a national conference [27], with an international audience, and also to an international conference [28].

It is a joint work with Laetitia Jourdan, from University of Lille.

7.29. Axis 4: Supervised multivariate discretization and levels merging for logistic regression

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

For regulatory and interpretability reasons, the logistic regression is still widely used by financial institutions to learn the refunding probability of a loan given the applicants characteristics from historical data. Although logistic regression handles naturally both quantitative and qualitative data, three ad hoc pre-processing steps are usually performed: firstly, continuous features are discretized by assigning factor levels to predetermined intervals; secondly, qualitative features, if they take numerous values, are grouped; thirdly, interactions (products between two different features) are sparsely introduced. By reinterpreting these discretized (resp. grouped) features as latent variables and by modeling the conditional distribution of each of these latent variables given each original feature with a polytomous logistic link (resp. contingency table), a novel model-based resolution of the discretization problem is introduced. Estimation is performed via a Stochastic Expectation-Maximization (SEM) algorithm and a Gibbs sampler to find the best discretization (resp. grouping) scheme w.r.t. any classical logistic regression loss (AIC, BIC, test set AUC, ...). For detecting interacting features, the same scheme is used by replacing the Gibbs sampler by a Metropolis-Hastings algorithm. The good performances of this approach are illustrated on simulated and real data from Credit Agricole Consumer Finance. This work has been presenting to an international conference in statistics [35] and has been also submitting to an international conference in machine learning. [42].

This is a joint work with Philippe Heinrich from University of Lille.

7.30. Axis 4: MASSICCC Platform for SaaS Software Availability

Participant: Christophe Biernacki.

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (forMixmod,MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2018, MASSICCC has been presented to a workshop [29]. Currently, a preprint for an international journal dedicated to software is also in progress.

The MASSICCC platform is available here in the web: https://massiccc.lille.inria.fr.

7.31. Axis 4: ClinMine: Optimizing the Management of Patients in Hospital

Participants: Cristian Preda, Vincent Vandewalle.

A better understanding of “patient pathway” thanks to data analysis can lead to better treatments for patients. The ClinMine project, supported by the French National Research Agency (ANR), aims at proposing, from various case studies, algorithmic and statistical models able to handle this type of pathway data, focusing primarily on hospital data.

Case studies, focusing on the integration of temporal data within analysis has been published [14]. First, the hypothesis that some aspects of the patient pathway can be described, even predicted, from the management process of the hospital medical mail is studied. Therefore a specific functional data analysis is driven, and several types of patients have been detected. The second case study deals with the detection of profiles through a biclustering of the patients. The difficulty to simultaneously deal with heterogeneous data, including temporal data is exposed and a method is proposed.

7.32. Projection Under Pairwise Control

Participant: Christophe Biernacki.

Visualization of high-dimensional and possibly complex (non-continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in R2 with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction. This work is still under revision in an international journal [48].

It is a joint work with Hiba Alawieh and Nicolas Wicker, both from University of Lille.

NON-A POST Team

6. New Results

6.1. Implementation of finite- and fixed-time algorithms

In [22] several algorithms of implicit discretization for generalized homogeneous systems having discontinuity only at the origin are developed. They are based on the transformation of the original system to an equivalent one which admits an implicit or a semi-implicit discretization schemes preserving the stability properties of the continuous-time system. Namely, the discretized model remains finite-time stable (in the case of negative homogeneity degree), and practically fixed-time stable (in the case of positive homogeneity degree).

6.2. A solution to finite- and fixed-time estimation

The work [18] deals with the problem of finite-time and fixed-time observation of linear multiple input multiple output (MIMO) control systems. The proposed nonlinear dynamic observers guarantee convergence of the observer states to the original system state in a finite and in a fixed (defined *a priori*) time. Algorithms for the observers parameters tuning are also provided and a robustness analysis against input disturbances and measurement noises is carried out.

6.3. Numeric and analytic design of homogeneous Lyapunov functions

The problem of the synthesis of a homogeneous Lyapunov function for an asymptotically stable homogeneous system is studied in [10]. First, for systems with nonnegative degree of homogeneity, several expressions of homogeneous Lyapunov functions are derived, which depend explicitly on the supremum or the integral (over finite or infinite intervals of time) of the system solutions. Second, a numeric procedure is proposed, which ensures the construction of a homogeneous Lyapunov function.

6.4. Distributed finite-time estimation

In [29] the robust distributed estimation for a class of time-invariant plants is achieved via a finite-time observer, its error reaching zero after a finite time in the absence of perturbation. Two types of robustness are also shown. First, input-to-state stability with respect to measurement noises and additive perturbations is proven. Second, we demonstrate that the estimation error stays bounded in the presence of known transmission delays.

RAPSODI Project-Team

7. New Results

7.1. Numerical simulation of concrete carbonation

In [20], C. Chainais-Hillairet, B. Merlet, and A. Zurek introduce and study a Finite Volume scheme for a concrete carbonation model proposed by Aiki and Muntean in [50]. This model consists in a system of two weakly coupled parabolic equations in a varying domain whose length is governed by an ordinary differential equation. The numerical scheme is obtained by a Euler discretization in time and a Scharfetter–Gummel discretization in space. The convergence of the scheme is established and the existence of a solution to the model is obtained as a by-product. Finally, some numerical experiments are performed to show the efficiency of the scheme.

In [45], A. Zurek studies the long-time regime of the moving interface appearing in the concrete carbonation model. He proves that the approximate free boundary, given by an implicit-in-time Finite Volume scheme, increases in time following a $\sqrt{t}$-law. This result is illustrated by numerical experiments.

7.2. Modeling and numerical simulation of complex fluids

In the context of C. Colin-Lecerf's PhD, C. Calgaro Zotto, C. Colin-Lecerf, and E. Creusé derive in [35] a combined Finite Volume-Finite Element scheme for a low-Mach model, in which a temperature field obeying an energy law is taken into account. The continuity equation is solved, whereas the state equation linking temperature, density, and thermodynamic pressure is imposed implicitly. Since the velocity field is not divergence-free, the projection method solving the momentum equation has to be adapted. This combined scheme preserves some steady states, and ensures a discrete maximum principle on the density. Numerical results are provided and compared to other approaches using purely Finite Element schemes, on a benchmark consisting in particular in a transient injection flow [58], [89], [53], as well as in the natural convection of a flow in a cavity [97], [93], [89], [53].

The theoretical study of the low-Mach limit system is a vast subject that has been considered by many authors. In particular, in [86], Embid establishes the local-in-time existence of classical solutions in Sobolev spaces. In [77], Danchin and Liao study the well-posedness issue in the critical Besov spaces, locally and globally, assuming that the initial density is close to a constant and that the initial velocity is small enough. Levermore *et al.* [98] consider the so-called ghost effect system, which is quite similar to the low-Mach system with thermal stress term added to the right-hand-side of the momentum equation, and they prove the local well-posedness of classical solutions for the Cauchy problem. In [94], Huang and Tan prove a local well-posedness result for strong solutions and also the existence and uniqueness of a global strong solution for the two-dimensional case. In [14], C. Calgaro Zotto, C. Colin-Lecerf, E. Creusé *et al.* investigate a specific low-Mach model for which the dynamic viscosity of the fluid is a specific function of the density. The model is reformulated in terms of the temperature and velocity, with nonlinear temperature equation, and strong solutions are considered. In addition to a local-in-time existence result for strong solutions, some convergence rates of the error between the approximation and the exact solution are obtained, following the same approach as Guillén-González *et al.* [91], [92].

Diffuse interface models, such as the Kazhikhov–Smagulov model, allow to describe some phase transition phenomena. In [15], C. Calgaro Zotto and co-workers investigate theoretically the combined Finite Volume-Finite Element scheme. They construct a fully discrete numerical scheme for approximating the two-dimensional Kazhikhov–Smagulov model, using a first-order time discretization and a splitting in time to allow the construction of the combined scheme. Consequently, at each time step, one only needs to solve two decoupled problems, the first one for the density (using the Finite Volume method) and the second one for the velocity and pressure (using the Finite Element method). The authors prove the stability of the combined scheme and the convergence towards the global-in-time weak solution of the model.

In [27], I. Lacroix-Violet *et al.* present the construction of global weak solutions to the quantum Navier–Stokes equation, for any initial value with bounded energy and entropy. The construction is uniform with respect to the Planck constant. This allows to perform the semi-classical limit to the associated compressible Navier–Stokes equation. One of the difficulties of the problem is to deal with the degenerate viscosity, together with the lack of integrability on the velocity. The method is based on the construction of weak solutions that are renormalized in the velocity variable. The existence and stability of these solutions do not need the Mellet–Vasseur inequality.

In [34], I. Lacroix-Violet *et al.* generalize to the Navier–Stokes–Korteweg (with density-dependent viscosities satisfying the BD relation) and Euler–Korteweg systems a recent relative entropy proposed in [65]. As a concrete application, this helps justifying mathematically the convergence between global weak solutions of the quantum Navier–Stokes system and dissipative solutions of the quantum Euler system when the viscosity coefficient tends to zero. The results are based on the fact that Euler–Korteweg systems and corresponding Navier–Stokes–Korteweg systems can be reformulated through an augmented system. As a by-product of the analysis, Lacroix-Violet *et al.* show that this augmented formulation helps to define relative entropy estimates for the Euler–Korteweg systems in a simpler way and with less hypotheses compared to recent works [82], [88].

7.3. Stratigraphic modeling and simulation

Stratigraphy is a discipline of physics that aims at predicting the geological composition of the subsoil. In [44], N. Peton, C. Cancès *et al.* propose a new water flow driven forward stratigraphic model with the following particularities. First, the water surface flow is modelled at the continuous level, in opposition to what is currently done in this community. Second, the model incorporates a constraint on the erosion rate. A stable numerical scheme is proposed to simulate the model.

7.4. Numerical simulation in low-frequency electromagnetism

In [24], [28], E. Creusé and co-workers investigate the behavior of some Finite Element error estimators in the context of low-frequency elecromagnetism simulations, to underline the main differences in some practical situations. In addition, a more theoretical contribution is developed in [23], to prove the equivalence of some usual discrete gauge conditions. Once again, their numerical behaviors are compared on some characteristic benchmarks.

7.5. Asymptotic analysis

In [33], C. Cancès and co-workers derive the porous medium equation as the hydrodynamic limit of an interacting particle system which belongs to the family of exclusion processes with nearest neighbor exchanges. The main outcome of this work is to allow regions with vanishing density, where the dynamics turns out to degenerate. The convergence builds on a generalization of the entropy method and on suitable regularization of the dynamics.

In [29], A. Ait Hammou Oulhaj, C. Cancès, C. Chainais-Hillairet *et al.* study analytically and numerically the large time behavior of the solutions to a two-phase extension of the porous medium equation, which models the so-called seawater intrusion problem. They identify the self-similar solutions that correspond to steady states of a rescaled version of the problem. They finally provide numerical illustrations of the stationary states and exhibit numerical convergence rates.

In [13], C. Chainais-Hillairet *et al.* propose a new proof of existence of a solution to the scheme introduced in [63] which does not require any assumption on the time step. The result relies on the application of a topological degree argument which is based on the positivity and on uniform-in-time upper bounds of the approximate densities. They also establish uniform-in-time lower bounds satisfied by the approximate densities. These uniform-in-time upper and lower bounds ensure the exponential decay of the scheme towards the thermal equilibrium as shown in [63].

In [38], C. Chainais-Hillairet and M. Herda study the large-time behavior of solutions to Finite Volume discretizations of convection-diffusion equations or systems endowed with non-homogeneous Dirichlet and Neumann type boundary conditions. Their results concern various linear and nonlinear models such as Fokker–Planck equations, porous media equations, or drift-diffusion systems for semiconductors. For all of these models, some relative entropy principle is satisfied and implies exponential decay to the stationary state. They show that in the framework of Finite Volume schemes on orthogonal meshes, a large class of two-point monotone fluxes preserve this exponential decay of the discrete solution to the discrete steady state of the scheme.

In [32], M. Herda, T. Rey *et al.* are interested in the asymptotic analysis of a Finite Volume scheme for one-dimensional linear kinetic equations, with either Fokker–Planck or linearized BGK collision operator. Thanks to appropriate uniform estimates, they establish that the proposed scheme is asymptotic-preserving in the diffusive limit. Moreover, they adapt to the discrete framework the hypocoercivity method proposed by [80] to prove the exponential return to equilibrium of the approximate solution. They obtain decay estimates that are uniform in the diffusive limit. Finally, they present an efficient implementation of the proposed numerical schemes, and perform numerous numerical simulations assessing their accuracy and efficiency in capturing the correct asymptotic behaviors of the models.

In [26], M. Herda *et al.* consider various sets of Vlasov–Fokker–Planck equations modeling the dynamics of charged particles in a plasma under the effect of a strong magnetic field. For each of them, in a regime where the strength of the magnetic field is effectively stronger than that of collisions, they first formally derive asymptotically reduced models. In this regime, strong anisotropic phenomena occur; while equilibrium along magnetic field lines is asymptotically reached the asymptotic models capture a nontrivial dynamics in the perpendicular directions. They do check that in any case the obtained asymptotic model defines a well-posed dynamical system and when self-consistent electric fields are neglected they provide a rigorous mathematical justification of the formally derived systems. In this last step they provide a complete control on solutions by developing anisotropic hypocoercive estimates.

7.6. Structure-preserving numerical methods

The design and the analysis of numerical methods preserving at the discrete level the key features of the continuous models is one of the core tasks of the RAPSODI project-team. C. Cancès was invited to write a review paper [16] on energy stable numerical methods for complex porous media flows. The paper addresses three different approaches: monotonicity-based numerical methods like two-point flux approximation Finite Volumes, as well as two methods based on multi-point flow approximation that are either based on upwinding or on positive local dissipation tensors.

Concerning methods based on upwinding, A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet extend in [12] the nonlinear Control Volume Finite Element scheme of [69] to the discretization of Richards equation modeling unsaturated flows in porous media. This strategy is also applied in [30] by A. Ait Hammou Oulhaj and D. Maltese for the simulation of seawater intrusion in the subsoil nearby coastal regions. The scheme proposed in [30] is still convergent if the porous medium is anisotropic, in opposition to the energy-diminishing scheme analyzed in [11] by A. Ait Hammou Oulhaj, which is designed to be accurate in the long-time regime studied in [29]. Besides, an implicit Euler-Finite Volume scheme for a degenerate cross-diffusion system describing the ion transport through biological membranes is analyzed in [17] by C. Cancès, C. Chainais-Hillairet *et al.* The strongly coupled equations for the ion concentrations include drift terms involving the electric potential, which is coupled to the concentrations through the Poisson equation. The cross-diffusion system possesses a formal gradient flow structure revealing nonstandard degeneracies, which lead to considerable mathematical difficulties. The Finite Volume scheme is based on two-point flux approximations with "double" upwind mobilities. It preserves the structure of the continuous model like non-negativity, upper bounds, and entropy dissipation.

Concerning methods based on positive local dissipation tensors, C. Cancès, C. Chainais-Hillairet *et al.* propose in [18] a nonlinear Discrete Duality Finite Volume scheme to approximate the solutions of drift diffusion equations. The scheme is built to preserve at the discrete level even on severely distorted meshes

the energy/energy dissipation relation. In [37], C. Cancès and co-workers propose a Finite Element scheme for the numerical approximation of degenerate parabolic problems in the form of a nonlinear anisotropic Fokker–Planck equation. The scheme is energy-stable, only involves physically motivated quantities in its definition, and is able to handle general unstructured grids. Its convergence is rigorously proven thanks to compactness arguments, under very general assumptions. Although the scheme is based on Lagrange Finite Elements of degree 1, it is locally conservative after a local post-processing giving rise to an equilibrated flux. This also allows to derive a guaranteed *a posteriori* error estimate for the approximate solution. Numerical experiments are presented in order to give evidence of a very good behavior of the proposed scheme in various situations involving strong anisotropy and drift terms.

C. Cancès *et al.* derive in [36] a model of degenerate Cahn–Hilliard type for the phase segregation in incompressible multiphase flows. The model is obtained as the Wasserstein gradient flow of a Ginzburg–Landau energy with the constraint that the sum of the volume fractions must stay equal to 1. The resulting model differs from the classical degenerate Cahn–Hilliard model (see [106], [85]) and is closely related to a model proposed by E and collaborators [84], [100]. Besides the derivation of the model, the convergence of a minimizing movement scheme is proven in [36]. The Wasserstein gradient flow structure of the PDE system governing multiphase flows in porous media has recently been highlighted in [68]. The model can thus be approximated by means of the minimizing movement (or JKO) scheme, that C. Cancès *et al.* solve in [19] thanks to the ALG2-JKO scheme proposed in [60]. The numerical results are compared to a classical upstream mobility Finite Volume scheme, for which strong stability properties can be established.

In [42], S. Lemaire builds a bridge between the Hybrid High-Order [78] and Virtual Element [59] methods, which are the two main new-generation approaches to the arbitrary-order approximation of PDEs on meshes with general, polytopal cells. The Virtual Element method writes in functional terms and is naturally conforming; at the opposite, the Hybrid High-Order method writes in algebraic terms and is naturally nonconforming. It has been remarked a few years ago that the Hybrid High-Order method can be viewed as a nonconforming version of the Virtual Element method. In [42], S. Lemaire ends up unifying the Hybrid High-Order and Virtual Element approaches by showing that the Virtual Element method can be reformulated as a (newborn) conforming Hybrid High-Order method. This parallel has interesting consequences: it allows important simplifications in the *a priori* analysis of Virtual Element methods, and sheds new light on the differences between conforming and nonconforming Virtual Element methods, in particular in terms of mesh assumptions.

In [31], I. Lacroix-Violet *et al.* are interested in the numerical integration in time of nonlinear Schrödinger equations using different methods preserving the energy or a discrete analog of it. In particular, they give a rigorous proof of the order of the relaxation method (presented in [62] for cubic nonlinearities) and they propose a generalized version that allows to deal with general power law nonlinearities. Numerical simulations for different physical models show the efficiency of these methods.

7.7. Cost reduction for numerical methods

In [22], S. Lemaire *et al.* design and analyze (in the periodic setting) nonconforming multiscale methods for highly oscillatory elliptic problems, which are applicable on coarse grids that may feature general polytopal cells. Two types of methods are introduced: a Finite Element-type method, that generalizes classical nonconforming multiscale Finite Element methods to general meshes and to arbitrary-order polynomial cell boundary conditions, and a Virtual Element-type method, that allows, up to the computation of an adequate projection, to compute less oscillatory basis functions for equivalent precision. The Virtual Element-type method is based on the Hybrid High-Order framework [78]. As standard with such multiscale approaches, the general workflow of the method splits into an offline, massively parallelizable stage, where all fine-scale computations are performed, and the online, fully-coarse-scale stage.

In [25], T. Rey *et al.* extend the Fast Kinetic Scheme (FKS) originally constructed for solving the BGK equation, to the more challenging case of the Boltzmann equation. The scheme combines a robust and fast method for treating the transport part based on an innovative Lagrangian technique, supplemented with conservative fast spectral schemes to treat the collisional operator by means of an operator splitting approach.

This approach along with several implementation features related to the parallelization of the algorithm permits to construct an efficient simulation tool which is numerically tested against exact and reference solutions on classical problems arising in rarefied gas dynamics.

In [43], T. Rey *et al.* present high-order, fully explicit time integrators for nonlinear collisional kinetic equations, including the full Boltzmann equation. The methods, called projective integration, first take a few small steps with a simple, explicit method (forward Euler) to damp out the stiff components of the solution. Then, the time derivative is estimated and used in a Runge–Kutta method of arbitrary order. The procedure can be recursively repeated on a hierarchy of projective levels to construct telescopic projective integration methods. We illustrate the method with numerical results in one and two spatial dimensions.

7.8. Applied calculus of variations

In [41], B. Merlet *et al.* study a variational problem which models the behavior of topological singularities on the surface of a biological membrane in P_β-phase (see [103]). The problem combines features of the Ginzburg–Landau model in 2D and of the Mumford–Shah functional. As in the classical Ginzburg–Landau theory, a prescribed number of point vortices appear in the moderate energy regime; the model allows for discontinuities, and the energy penalizes their length. The novel phenomenon here is that the vortices have a fractional degree $1/m$ with m prescribed. Those vortices must be connected by line discontinuities to form clusters of total integer degrees. The vortices and line discontinuities are therefore coupled through a topological constraint. As in the Ginzburg–Landau model, the energy is parameterized by a small length scale $\varepsilon > 0$. B. Merlet *et al.* perform a complete Γ-convergence analysis of the model as $\varepsilon \downarrow 0$ in the moderate energy regime. Then, they study the structure of minimizers of the limit problem. In particular, the line discontinuities of a minimizer solve a variant of the Steiner problem.

In [21], B. Merlet *et al.* consider a generalization of branched transport in arbitrary dimension and codimension: minimize the h-mass of some oriented k-dimensional branched surface in $\mathbf{R}^n$ with some prescribed boundary. Attached to the surface is a multiplicity $m(x)$ which is not necessarily an integer and is a conserved quantity (Kirchhoff current law is satisfied at branched points). The h-mass is defined as the integral of a cost $h(|m(x)|)$ over the branched surface. As usual in branched transportation, the cost function is a lower-semicontious, sublinear increasing function with $h(0) = 0$ (for instance $h(m) = \sqrt{1 + am^2}$ if $m \neq 0$ and $h(0) = 0$). For numerical purpose, it is convenient to approximate the measure defined by the k-dimensional surfaces by smooth functions in $\mathbf{R}^n$. In this spirit, B. Merlet *et al.* propose phase field approximations of the branched surfaces and of their energy in the spirit of the Ambrosio–Tortorelli functional. The convergence of these approximations towards the original k-dimensional branched transport problem is established in [21] in the sense of Γ-convergence. Next, considering the cost $h(m) = \sqrt{1 + am^2}$ and sending a to 0, a phase field approximation of the Plateau problem is obtained. Numerical experiments show the efficiency of the method. These numerical results are exceptional as they are obtained without any guess on the topology of the minimizing k-surface (as opposed to methods based on parameterizations of the k-surface). In [39], B. Merlet *et al.* establish new results on the approximation of k-dimensional surfaces (k-rectifiable currents) by polyhedral surfaces with convergence in h-mass and with preservation of the boundary (the approximating polyhedral surface has the same boundary as the limit). This approximation result is required in the convergence study of [21].

7.9. Approximation theory

In [40], M. Herda *et al.* propose a new iterative algorithm for the calculation of sum of squares decompositions of polynomials, reformulated as positive interpolation. The method is based on the definition of a dual functional G from values at interpolation points. The domain of G, the boundary of the domain and the behavior of G at infinity are analyzed in details. In the general case, G is closed convex. For univariate polynomials in the context of the Lukacs representation, G is coercive and strictly convex which yields a unique critical point, corresponding to a sum of squares decomposition of G. Various descent algorithms are evoked. Numerical examples are provided, for univariate and bivariate polynomials.

RMOD Project-Team

7. New Results

7.1. Dynamic Languages: Virtual Machines

Assessing primitives performance on multi-stage execution. Virtual machines, besides the interpreter and just-in-time compiler optimization facilities, also include a set of primitive operations that the client language can use. Some of these are essential and cannot be performed in any other way. Others are optional: they can be expressed in the client language but are often implemented in the virtual machine to improve performance when the just-in-time compiler is unable to do so (start-up performance, speculative optimizations not implemented or not mature enough, etc.). In a hybrid runtime, where code is executed by an interpreter and a just-in-time compiler, the implementor can choose to implement optional primitives in the client language , in the virtual machine implementation language (typically C or C++), or on top of the just-in-time compiler back-end. This raises the question of the maintenance and performance trade-offs of the different alternatives. As a case study, we implemented the String comparison optional primitive in each case. The paper describes the different implementations, discusses the maintenance cost of each of them and evaluates for different string sizes the execution time in Cog, a Smalltalk virtual machine. [18]

Fully Reflective Execution Environments: Virtual Machines for More Flexible Software. VMs are complex pieces of software that implement programming language semantics in an efficient, portable, and secure way. Unfortunately, mainstream VMs provide applications with few mechanisms to alter execution semantics or memory management at run time. We argue that this limits the evolvability and maintainability of running systems for both, the application domain, e.g., to support unforeseen requirements, and the VM domain, e.g., to modify the organization of objects in memory. This work explores the idea of incorporating reflective capabilities into the VM domain and analyzes its impact in the context of software adaptation tasks. We characterize the notion of a fully reflective VM, a kind of VM that provides means for its own observability and modifiability at run time. This enables programming languages to adapt the underlying VM to changing requirements. We propose a reference architecture for such VMs and present TruffleMATE as a prototype for this architecture. We evaluate the mechanisms TruffleMATE provides to deal with unanticipated dynamic adaptation scenarios for security, optimization, and profiling aspects. In contrast to existing alternatives, we observe that TruffleMATE is able to handle all scenarios, using less than 50 lines of code for each, and without interfering with the application's logic. [2]

7.2. Dynamic Languages: Language Constructs for Modular Design

Dynamic Software Update from Development to Production. Dynamic Software Update (DSU) solutions update applications while they are executing. These solutions are typically used in production to minimize application downtime, or in integrated development environments to provide live programming support. Each of these scenarios presents different challenges, forcing existing solutions to be designed with only one of these use cases in mind. For example, DSUs for live programming typically do not implement safe point detection or instance migration, while production DSUs require manual generation of patches and lack IDE integration. Also, these solutions have limited ability to update themselves or the language core libraries, and some of them present execution penalties outside the update window. We propose a DSU (gDSU) that works for both live programming and production environments. Our solution implements safe update point detection using call stack manipulation and a reusable instance migration mechanism to minimize manual intervention in patch generation. Moreover, it also offers updates of core language libraries and the update mechanism itself. This is achieved by the incremental copy of the modified objects and an atomic commit operation. We show that our solution does not affect the global performance of the application and it presents only a run-time penalty during the update window. Our solution is able to apply an update impacting 100,000 instances in 1 second. In this 1 second, only during 250 milliseconds the application is not responsive. The rest of the time the application runs normally while gDSU is looking for the safe update point. The update only requires to copy the elements that are modified. [6]

Implementing Modular Class-based Reuse Mechanisms on Top of a Single Inheritance VM. Code reuse is a good strategy to avoid code duplication and speed up software development. Existing object-oriented programming languages propose different ways of combining existing and new code such as e.g., single inheritance, multiple inheritance, Traits or Mixins. All these mechanisms present advantages and disadvantages and there are situations that require the use of one over the other. To avoid the complexity of implementing a virtual machine (VM), many of these mechanisms are often implemented on top of an existing high-performance VM, originally meant to run a single inheritance object-oriented language. These implementations require thus a mapping between the programming model they propose and the execution model provided by the VM. Moreover, reuse mechanisms are not usually composable, nor it is easy to implement new ones for a given language. We propose a modular meta-level runtime architecture to implement and combine different code reuse mechanisms. This architecture supports dynamic combination of several mechanisms without affecting runtime performance in a single inheritance object-oriented VM. It includes moreover a reflective Meta-Object Protocol to query and modify classes using the programming logical model instead of the underlying low-level runtime model. Thanks to this architecture, we implemented Stateful Traits, Mixins, CLOS multiple inheritance, CLOS Standard Method Combinations and Beta prefixing in a modular and composable way. [15]

7.3. Software Reengineering

A Reflexive and Automated Approach to Syntactic Pattern Matching in Code Transformations. Empowering software engineers often requires to let them write code transformations. However existing automated or tool-supported approaches force developers to have a detailed knowledge of the internal representation of the underlying tool. While this knowledge is time consuming to master, the syntax of the language, on the other hand, is already well known to developers and can serve as a strong foundation for pattern matching. Pattern languages with metavariables (that is variables holding abstract syntax subtrees once the pattern has been matched) have been used to help programmers define program transformations at the language syntax level. The question raised is then the engineering cost of metavariable support. Our contribution is to show that, with a GLR parser, such patterns with metavariables can be supported by using a form of runtime reflexivity on the parser internal structures. This approach allows one to directly implement such patterns on any parser generated by a parser generation framework, without asking the pattern writer to learn the AST structure and node types. As a use case for that approach we describe the implementation built on top of the SmaCC (Smalltalk Compiler Compiler) GLR parser generator framework. This approach has been used in production for source code transformations on a large scale. We will express perspectives to adapt this approach to other types of parsing technologies. [12]

Relational Database Schema Evolution: An Industrial Case Study. Modern relational database management systems provide advanced features allowing, for example, to include behavior directly inside the database (stored procedures). These features raise new difficulties when a database needs to evolve (e.g. adding a new table). To get a better understanding of these difficulties, we recorded and studied the actions of a database architect during a complex evolution of the database at the core of a software system. From our analysis, problems faced by the database architect are extracted, generalized and explored through the prism of software engineering. Six problems are identified: (1) difficulty in analyzing and visualizing dependencies between database's entities, (2) difficulty in evaluating the impact of a modification on the database, (3) replicating the evolution of the database schema on other instances of the database, (4) difficulty in testing database's functionalities, (5) lack of synchronization between the IDE's internal model of the database and the database actual state and (6) absence of an integrated tool enabling the architect to search for dependencies between entities, generate a patch or access up to date PostgreSQL documentation. We suggest that techniques developed by the software engineering community could be adapted to help in the development and evolution of relational databases. [10]

A Quality-oriented Approach to Recommend Move Method Refactorings. Refactoring is an important activity to improve software internal structure. Even though there are many refactoring approaches, very few consider their impact on the software quality. We propose a software refactoring approach based on quality

attributes. We rely on the measurements of the Quality Model for Object Oriented Design (QMOOD) to recommend Move Method refactorings that improve software quality. In a nutshell, given a software system S, our approach recommends a sequence of refactorings R1,R2,...,Rn that result in system versions S1, S2,..., Sn, where quality (Si+1) > quality (Si). We empirically calibrated our approach, using four systems, to find the best criteria to measure the quality improvement. We performed three types of evaluation to verify the usefulness of our implemented tool, named QMove. First, we applied our approach on 13 open-source systems achieving an average recall of 84.2%. Second, we compared QMove with two state-of-art refactoring tools (JMove and JDeodorant) on the 13 previously evaluated systems, and QMove showed better recall, precision, and f-score values than the others. Third, we evaluated QMove, JMove, and JDeodorant in a real scenario with two proprietary systems on the eyes of their software architects. As result, the experts positively evaluated a greater number of QMove recommendations. [14]

7.4. Dynamic Languages: Debugging

Collectors. Observing and modifying object-oriented programs often means interacting with objects. At runtime, it can be a complex task to identify those objects due to the live state of the program. Some objects may exist for only a very limited period of time, others can be hardly reachable because they are never stored in variables. To address this problem we present Collectors. They are dedicated objects which can collect objects of interest at runtime and present them to the developer. Collectors are non-intrusive, removable code instrumentations. They can be dynamically specified and injected at runtime. They expose an API to allow their specification and the access to the collected objects. We present an implementation of Collectors in Pharo, a Smalltalk dialect. We enrich the Pharo programming and debugging environment with tools that support the Collectors API. We illustrate the use of these API and tools through the collection and the logging of specific objects in a running IOT application. [9]

Rotten Green Tests: a First Analysis. Unit tests are a tenant of agile programming methodologies, and are widely used to improve code quality and prevent code regression. A passing (green) test is usually taken as a robust sign that the code under test is valid. However, we have noticed that some green tests contain assertions that are never executed; these tests pass not because they assert properties that are true, but because they assert nothing at all. We call such tests Rotten Green Tests. Rotten Green Tests represent a worst case: they report that the code under test is valid, but in fact do nothing to test that validity, beyond checking that the code does not crash. We describe an approach to identify rotten green tests by combining simple static and dynamic analyses. Our approach takes into account test helper methods, inherited helpers, and trait compositions, and has been implemented in a tool called DrTest. We have applied DrTest to several test suites in Pharo 7.0, and identified many rotten tests, including some that have been sleeping in Pharo for at least 5 years. [22]

Mining inline cache data to order inferred types in dynamic languages. The lack of static type information in dynamically-typed languages often poses obstacles for developers. Type inference algorithms can help, but inferring precise type information requires complex algorithms that are often slow. A simple approach that considers only the locally used interface of variables can identify potential classes for variables, but popular interfaces can generate a large number of false positives. We propose an approach called inline-cache type inference (ICTI) to augment the precision of fast and simple type inference algorithms. ICTI uses type information available in the inline caches during multiple software runs, to provide a ranked list of possible classes that most likely represent a variable's type. We evaluate ICTI through a proof-of-concept that we implement in Pharo Smalltalk. The analysis of the top-n+2 inferred types (where n is the number of recorded run-time types for a variable) for 5486 variables from four different software systems shows that ICTI produces promising results for about 75% of the variables. For more than 90% of variables, the correct run-time type is present among first six inferred types. Our ordering shows a twofold improvement when compared with the unordered basic approach, i.e., for a significant number of variables for which the basic approach offered ambiguous results, ICTI was able to promote the correct type to the top of the list. [22]

7.5. Blockchain

Ethereum Query Language Blockchains store a massive amount of heterogeneous data which will only grow in time. When searching for data on the Ethereum platform, one is required to either access the records (blocks) directly by using a unique identifier, or sequentially search several records to find the desired information. Therefore, we propose the Ethereum Query Language (EQL), a query language that allows users to retrieve information from the blockchain by writing SQL-like queries. The queries provide a rich syntax to specify data elements to search information scattered through several records. We claim that EQL makes it easier to search, acquire, format, and present information from the blockchain. [7]

SmartInspect: solidity smart contract inspector. Solidity is a language used for smart contracts on the Ethereum blockchain. Smart contracts are embedded procedures stored with the data they act upon. Debugging smart contracts is a really difficult task since once deployed, the code cannot be reexecuted and inspecting a simple attribute is not easily possible because data is encoded. We address the lack of inspectability of a deployed contract by analyzing contract state using decompilation techniques driven by the contract structure definition. Our solution, SmartInspect, also uses a mirror-based architecture to represent locally object responsible for the interpretation of the contract state. SmartInspect allows contract developers to better visualize and understand the contract stored state without needing to redeploy, nor develop any ad-hoc code. [8]

Preliminary Steps Towards Modeling Blockchain Oriented Software Even though blockchain is mostly popular for its cryptocurrency, smart contracts have become a very prominent blockchain application. Smart contracts are like classes that can be called by client applications outside the blockchain. Therefore it is possible to develop blockchain-oriented software (BOS) that implements part of the business logic in the blockchain by using smart contracts. Currently, there is no design standard to model BOS. Since modeling is an important part of designing a software, developers may struggle to plan their BOS. We show three complementary modeling approaches based on well-known software engineering models and apply them to a BOS example. Our goal is to start the discussion on specialized blockchain modeling notations. [13]

SmartAnvil: Open-Source Tool Suite for Smart Contract Analysis. Smart contracts are new computational units with special properties: they act as classes with aspectual concerns; their memory structure is more complex than mere objects; they are obscure in the sense that once deployed it is difficult to access their internal state; they reside in an append-only chain. There is a need to support the building of new generation tools to help developers. Such support should tackle several important aspects: (1) the static structure of the contract, (2) the object nature of published contracts, and (3) the overall data chain composed of blocks and transactions. In this chapter, we present SmartAnvil an open platform to build software analysis tools around smart contracts. We illustrate the general components and we focus on three important aspects: support for static analysis of Solidity smart contracts, deployed smart contract binary analysis through inspection, and blockchain navigation and querying. SmartAnvil is open-source and supports a bridge to the Moose data and software analysis platform. [21]

SEQUEL Project-Team

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

A Fitted-Q Algorithm for Budgeted MDPs, [26]

We address the problem of budgeted reinforcement learning, in continuous state-space, using a batch of transitions. To this extend, we introduce a novel algorithm called Budgeted Fitted-Q (BFTQ). Benchmarks show that BFTQ performs as well as a regular Fitted-Q algorithm in a continuous 2-D world but also allows one to choose the right amount of budget that fits to a given task without the need of engineering the rewards. We believe that the general principles used to design BFTQ can be applied to extend others classical reinforcement learning algorithms for budgeted oriented applications.

Safe transfer learning for dialogue applications, [27]

In this paper, we formulate the hypothesis that the first dialogues with a new user should be handle in a very conservative way, for two reasons : avoid user dropout; gather more successful dialogues to speedup the learning of the asymptotic strategy. To this extend, we propose to transfer a safe strategy to initiate the first dialogues.

Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs, [17]

The problem of reinforcement learning in an unknown and discrete Markov Decision Process (MDP) under the average-reward criterion is considered, when the learner interacts with the system in a single stream of observations, starting from an initial state without any reset. We revisit the minimax lower bound for that problem by making appear the local variance of the bias function in place of the diameter of the MDP. Furthermore, we provide a novel analysis of the KL-Ucrl algorithm establishing a high-probability regret bound scaling as O S s,a V s,a T for this algorithm for ergodic MDPs, where S denotes the number of states and where V s,a is the variance of the bias function with respect to the next-state distribution following action a in state s. The resulting bound improves upon the best previously known regret bound O(DS $\sqrt{}$ AT) for that algorithm, where A and D respectively denote the maximum number of actions (per state) and the diameter of MDP. We finally compare the leading terms of the two bounds in some benchmark MDPs indicating that the derived bound can provide an order of magnitude improvement in some cases. Our analysis leverages novel variations of the transportation lemma combined with Kullback-Leibler concentration inequalities, that we believe to be of independent interest.

Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning, [29]

We introduce SCAL, an algorithm designed to perform efficient exploration-exploitation in any unknown weakly-communicating Markov decision process (MDP) for which an upper bound c on the span of the optimal bias function is known. For an MDP with S states, A actions and $\Gamma \leq S$ possible next states, we prove a regret bound of $\widetilde{O}(c\sqrt{\Gamma SAT})$, which significantly improves over existing algorithms (e.g., UCRL and PSRL), whose regret scales linearly with the MDP diameter D. In fact, the optimal bias span is finite and often much smaller than D (e.g., $D = \infty$ in non-communicating MDPs). A similar result was originally derived by Bartlett and Tewari (2009) for REGAL.C, for which no tractable algorithm is available. In this paper, we relax the optimization problem at the core of REGAL.C, we carefully analyze its properties, and we provide the first computationally efficient algorithm to solve it. Finally, we report numerical simulations supporting our theoretical findings and showing how SCAL significantly outperforms UCRL in MDPs with large diameter and small span.

Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes, [28]

While designing the state space of an MDP, it is common to include states that are transient or not reachable by any policy (e.g., in mountain car, the product space of speed and position contains configurations that are not physically reachable). This leads to defining weakly-communicating or multi-chain MDPs. In this paper, we introduce TUCRL, the first algorithm able to perform efficient exploration-exploitation in any finite Markov Decision Process (MDP) without requiring any form of prior knowledge. In particular, for any MDP with S^{C} communicating states, A actions and $\Gamma^{\mathsf{C}} \leq S^{\mathsf{C}}$ possible communicating next states, we derive a $\widetilde{O}(D^{\mathsf{C}}\sqrt{\Gamma^{\mathsf{C}} S^{\mathsf{C}} AT})$ regret bound, where D^{C} is the diameter (i.e., the longest shortest path) of the communicating part of the MDP. This is in contrast with optimistic algorithms (e.g., UCRL, Optimistic PSRL) that suffer linear regret in weakly-communicating MDPs, as well as posterior sampling or regularized algorithms (e.g., REGAL), which require prior knowledge on the bias span of the optimal policy to bias the exploration to achieve sub-linear regret. We also prove that in weakly-communicating MDPs, no algorithm can ever achieve a logarithmic growth of the regret without first suffering a linear regret for a number of steps that is exponential in the parameters of the MDP. Finally, we report numerical simulations supporting our theoretical findings and showing how TUCRL overcomes the limitations of the state-of-the-art.

Upper Confidence Reinforcement Learning exploiting state-action equivalence, [53]

Stochastic Variance-Reduced Policy Gradient, [34]

In this paper, we propose a novel reinforcement-learning algorithm consisting in a stochastic variance-reduced version of policy gradient for solving Markov Decision Processes (MDPs). Stochastic variance-reduced gradient (SVRG) methods have proven to be very successful in supervised learning. However, their adaptation to policy gradient is not straightforward and needs to account for I) a non-concave objective function; II) approximations in the full gradient computation; and III) a non-stationary sampling process. The result is SVRPG, a stochastic variance-reduced policy gradient algorithm that leverages on importance weights to preserve the unbiasedness of the gradient estimate. Under standard assumptions on the MDP, we provide convergence guarantees for SVRPG with a convergence rate that is linear under increasing batch sizes. Finally, we suggest practical variants of SVRPG, and we empirically evaluate them on continuous MDPs.

Importance Weighted Transfer of Samples in Reinforcement Learning, [38]

We consider the transfer of experience samples (i.e., tuples < s, a, s', r >) in reinforcement learning (RL), collected from a set of source tasks to improve the learning process in a given target task. Most of the related approaches focus on selecting the most relevant source samples for solving the target task, but then all the transferred samples are used without considering anymore the discrepancies between the task models. In this paper, we propose a model-based technique that automatically estimates the relevance (importance weight) of each source sample for solving the target task. In the proposed approach, all the samples are transferred and used by a batch RL algorithm to solve the target task, but their contribution to the learning process is proportional to their importance weight. By extending the results for importance weighting provided in supervised learning literature, we develop a finite-sample analysis of the proposed batch RL algorithm. Furthermore, we empirically compare the proposed algorithm to state-of-the-art approaches, showing that it achieves better learning performance and is very robust to negative transfer, even when some source tasks are significantly different from the target task.

Training Dialogue Systems With Human Advice, [20]

One major drawback of Reinforcement Learning (RL) Spoken Dialogue Systems is that they inherit from the general exploration requirements of RL which makes them hard to deploy from an industry perspective. On the other hand, industrial systems rely on human expertise and hand written rules so as to avoid irrelevant behavior to happen and maintain acceptable experience from the user point of view. In this paper, we attempt to bridge the gap between those two worlds by providing an easy way to incorporate all kinds of human expertise in the training phase of a Reinforcement Learning Dialogue System. Our approach, based on the TAMER framework, enables safe and efficient policy learning by combining the traditional Reinforcement Learning reward signal with an additional reward, encoding expert advice. Experimental results show that our method leads to substantial improvements over more traditional Reinforcement Learning methods.

7.1.1.1. Deep reinforcement learning

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

Feature-wise transformations, [13]

Deep Reinforcement Learning and the Deadly Triad, [55]

We know from reinforcement learning theory that temporal difference learning can fail in certain cases. Sutton and Barto (2018) identify a deadly triad of function approximation, bootstrapping, and off-policy learning. When these three properties are combined, learning can diverge with the value estimates becoming unbounded. However, several algorithms successfully combine these three properties, which indicates that there is at least a partial gap in our understanding. In this work, we investigate the impact of the deadly triad in practice, in the context of a family of popular deep reinforcement learning models - deep Q-networks trained with experience replay - analyzing how the components of this system play a role in the emergence of the deadly triad, and in the agent's performance

7.1.2. Multi-armed Bandit Theory

Corrupt Bandits for Preserving Local Privacy, [30]

We study a variant of the stochastic multi-armed bandit (MAB) problem in which the rewards are corrupted. In this framework, motivated by privacy preservation in online recommender systems, the goal is to maximize the sum of the (unobserved) rewards, based on the observation of transformation of these rewards through a stochastic corruption process with known parameters. We provide a lower bound on the expected regret of any bandit algorithm in this corrupted setting. We devise a frequentist algorithm, KLUCB-CF, and a Bayesian algorithm, TS-CF and give upper bounds on their regret. We also provide the appropriate corruption parameters to guarantee a desired level of local privacy and analyze how this impacts the regret. Finally, we present some experimental results that confirm our analysis.

A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption, [21]

We study the problem of optimizing a function under a budgeted number of evaluations. We only assume that the function is locally smooth around one of its global optima. The difficulty of optimization is measured in terms of 1) the amount of noise b of the function evaluation and 2) the local smoothness, d, of the function. A smaller d results in smaller optimization error. We come with a new, simple, and parameter-free approach. First, for all values of b and d, this approach recovers at least the state-of-the-art regret guarantees. Second, our approach additionally obtains these results while being agnostic to the values of both b and d. This leads to the first algorithm that naturally adapts to an unknown range of noise b and leads to significant improvements in a moderate and low-noise regime. Third, our approach also obtains a remarkable improvement over the state-of-the-art SOO algorithm when the noise is very low which includes the case of optimization under deterministic feedback ($b = 0$). There, under our minimal local smoothness assumption, this improvement is of exponential magnitude and holds for a class of functions that covers the vast majority of functions that practitioners optimize ($d = 0$). We show that our algorithmic improvement is also borne out in the numerical experiments, where we empirically show faster convergence on common benchmark functions.

Best of both worlds: Stochastic & adversarial best-arm identification, [18]

We study bandit best-arm identification with arbitrary and potentially adversarial rewards. A simple random uniform learner obtains the optimal rate of error in the adversarial scenario. However, this type of strategy is suboptimal when the rewards are sampled stochastically. Therefore, we ask: Can we design a learner that performs optimally in both the stochastic and adversarial problems while not being aware of the nature of the rewards? First, we show that designing such a learner is impossible in general. In particular, to be robust to adversarial rewards, we can only guarantee optimal rates of error on a subset of the stochastic problems. We give a lower bound that characterizes the optimal rate in stochastic problems if the strategy is constrained to be robust to adversarial rewards. Finally, we design a simple parameter-free algorithm and show that its probability of error matches (up to log factors) the lower bound in stochastic problems, and it is also robust to adversarial ones.

Optimistic optimization of a Brownian, [31]

We address the problem of optimizing a Brownian motion. We consider a (random) realization W of a Brownian motion with input space in $[0, 1]$. Given W, our goal is to return an ϵ-approximation of its maximum using the smallest possible number of function evaluations, the sample complexity of the algorithm. We provide an algorithm with sample complexity of order $\log 2\ (1/\epsilon)$. This improves over previous results of Al-Mharmah and Calvin (1996) and Calvin et al. (2017) which provided only polynomial rates. Our algorithm is adaptive-each query depends on previous values-and is an instance of the optimism-in-the-face-of-uncertainty principle.

Rotting bandits are no harder than stochastic ones, [37]

In bandits, arms' distributions are stationary. This is often violated in practice, where rewards change over time. In applications as recommendation systems, online advertising, and crowdsourcing, the changes may be triggered by the pulls, so that the arms' rewards change as a function of the number of pulls. In this paper, we consider the specific case of non-parametric rotting bandits, where the expected reward of an arm may decrease every time it is pulled. We introduce the filtering on expanding window average (FEWA) algorithm that at each round constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. We prove that, without any knowledge on the decreasing behavior of the arms, FEWA achieves similar anytime problem-dependent, $\widetilde{\mathcal{O}}(\log(KT))$, and problem-independent, $\widetilde{\mathcal{O}}(\sqrt{KT})$, regret bounds of near-optimal stochastic algorithms as UCB1 of Auer et al. (2002a). This result substantially improves the prior result of Levine et al. (2017) which needed knowledge of the horizon and decaying parameters to achieve problem-independent bound of only $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$. Finally, we report simulations confirming the theoretical improvements of FEWA.

Adaptive black-box optimization got easier: HCT only needs local smoothness, [41]

Hierarchical bandits is an approach for global optimization of extremely irregular functions. This paper provides new elements regarding POO, an adaptive meta-algorithm that does not require the knowledge of local smoothness of the target function. We first highlight the fact that the subroutine algorithm used in POO should have a small regret under the assumption of local smoothness with respect to the chosen partitioning, which is unknown if it is satisfied by the standard subroutine HOO. In this work, we establish such regret guarantee for HCT, which is another hierarchical optimistic optimization algorithm that needs to know the smoothness. This confirms the validity of POO. We show that POO can be used with HCT as a subroutine with a regret upper bound that matches the one of best-known algorithms using the knowledge of smoothness up to a $\sqrt{}$ log n factor.

Boundary Crossing Probabilities for General Exponential Families, [16]

Multi-Player Bandits Revisited, [22]

Multi-player Multi-Armed Bandits (MAB) have been extensively studied in the literature, motivated by applications to Cognitive Radio systems. Driven by such applications as well, we motivate the introduction of several levels of feedback for multi-player MAB algorithms. Most existing work assume that sensing information is available to the algorithm. Under this assumption, we improve the state-of-the-art lower bound for the regret of any decentralized algorithms and introduce two algorithms, RandTopM and MCTopM, that are

shown to empirically outperform existing algorithms. Moreover, we provide strong theoretical guarantees for these algorithms, including a notion of asymptotic optimality in terms of the number of selections of bad arms. We then introduce a promising heuristic, called Selfish, that can operate without sensing information, which is crucial for emerging applications to Internet of Things networks. We investigate the empirical performance of this algorithm and provide some first theoretical elements for the understanding of its behavior.

Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence, [19]

We consider the problem of near-optimal arm identification in the fixed confidence setting of the infinitely armed bandit problem when nothing is known about the arm reservoir distribution. We (1) introduce a PAC-like framework within which to derive and cast results; (2) derive a sample complexity lower bound for near-optimal arm identification; (3) propose an algorithm that identifies a nearly-optimal arm with high probability and derive an upper bound on its sample complexity which is within a log factor of our lower bound; and (4) discuss whether our $\log^2(1/\delta)$ dependence is inescapable for "two-phase" (select arms first, identify the best later) algorithms in the infinite setting. This work permits the application of bandit models to a broader class of problems where fewer assumptions hold.

Aggregation of Multi-Armed Bandits Learning Algorithms for Opportunistic Spectrum Access, [23]

Multi-armed bandit algorithms have been recently studied and evaluated for Cognitive Radio (CR), especially in the context of Opportunistic Spectrum Access (OSA). Several solutions have been explored based on various models, but it is hard to exactly predict which could be the best for real-world conditions at every instants. Hence, expert aggregation algorithms can be useful to select on the run the best algorithm for a specific situation. Aggregation algorithms, such as Exp4 dating back from 2002, have never been used for OSA learning, and we show that it appears empirically sub-efficient when applied to simple stochastic problems. In this article, we present an improved variant, called Aggregator. For synthetic OSA problems modeled as Multi-Armed Bandit (MAB) problems, simulation results are presented to demonstrate its empirical efficiency. We combine classical algorithms, such as Thompson sampling, Upper-Confidence Bounds algorithms (UCB and variants), and Bayesian or Kullback-Leibler UCB. Our algorithm offers good performance compared to state-of-the-art algorithms (Exp4, CORRAL or LearnExp), and appears as a robust approach to select on the run the best algorithm for any stochastic MAB problem, being more realistic to real-world radio settings than any tuning-based approach.

What Doubling Tricks Can and Can't Do for Multi-Armed Bandits, [47]

An online reinforcement learning algorithm is anytime if it does not need to know in advance the horizon T of the experiment. A well-known technique to obtain an anytime algorithm from any non-anytime algorithm is the "Doubling Trick". In the context of adversarial or stochastic multi-armed bandits, the performance of an algorithm is measured by its regret, and we study two families of sequences of growing horizons (geometric and exponential) to generalize previously known results that certain doubling tricks can be used to conserve certain regret bounds. In a broad setting, we prove that a geometric doubling trick can be used to conserve (minimax) bounds in $R\_T = O(\sqrt{T})$ but cannot conserve (distribution-dependent) bounds in $R\_T = O(\log T)$. We give insights as to why exponential doubling tricks may be better, as they conserve bounds in $R\_T = O(\log T)$, and are close to conserving bounds in $R\_T = O(\sqrt{T})$.

Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals, [50]

This paper presents new deviation inequalities that are valid uniformly in time under adaptive sampling in a multi-armed bandit model. The deviations are measured using the Kullback-Leibler divergence in a given one-dimensional exponential family, and may take into account several arms at a time. They are obtained by constructing for each arm a mixture martingale based on a hierarchical prior, and by multiplying those martingales. Our deviation inequalities allow us to analyze stopping rules based on generalized likelihood ratios for a large class of sequential identification problems, and to construct tight confidence intervals for some functions of the means of the arms.

7.1.3. Stochastic Games

Actor-Critic Fictitious Play in Simultaneous Move Multistage Games, [36]

Fictitious play is a game theoretic iterative procedure meant to learn an equilibrium in normal form games. However, this algorithm requires that each player has full knowledge of other players' strategies. Using an architecture inspired by actor-critic algorithms, we build a stochastic approximation of the fictitious play process. This procedure is on-line, decentralized (an agent has no information of others' strategies and rewards) and applies to multistage games (a generalization of normal form games). In addition, we prove convergence of our method towards a Nash equilibrium in both the cases of zero-sum two-player multistage games and cooperative multistage games. We also provide empirical evidence of the soundness of our approach on the game of Alesia with and without function approximation.

Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling, [39]

Learning the minimum/maximum mean among a finite set of distributions is a fundamental sub-task in planning, game tree search and reinforcement learning. We formalize this learning task as the problem of sequentially testing how the minimum mean among a finite set of distributions compares to a given threshold. We develop refined non-asymptotic lower bounds, which show that optimality mandates very different sampling behavior for a low vs high true minimum. We show that Thompson Sampling and the intuitive Lower Confidence Bounds policy each nail only one of these cases. We develop a novel approach that we call Murphy Sampling. Even though it entertains exclusively low true minima, we prove that MS is optimal for both possibilities. We then design advanced self-normalized deviation inequalities, fueling more aggressive stopping rules. We complement our theoretical guarantees by experiments showing that MS works best in practice.

7.1.4. Online Kernel and Graph-Based Methods

Improved large-scale graph learning through ridge spectral sparsification, [25]

The representation and learning benefits of methods based on graph Laplacians, such as Laplacian smoothing or harmonic function solution for semi-supervised learning (SSL), are empirically and theoretically well supported. Nonetheless , the exact versions of these methods scale poorly with the number of nodes n of the graph. In this paper, we combine a spectral sparsification routine with Laplacian learning. Given a graph G as input, our algorithm computes a sparsifier in a distributed way in $O(n \log 3(n))$ time, $O(m \log 3(n))$ work and $O(n \log(n))$ memory, using only $\log(n)$ rounds of communication. Furthermore , motivated by the regularization often employed in learning algorithms, we show that constructing sparsifiers that preserve the spectrum of the Laplacian only up to the regularization level may drastically reduce the size of the final graph. By constructing a spectrally-similar graph, we are able to bound the error induced by the sparsification for a variety of downstream tasks (e.g., SSL). We empirically validate the theoretical guarantees on Amazon co-purchase graph and compare to the state-of-the-art heuristics.

DPPy: Sampling Determinantal Point Processes with Python, [49]

Determinantal point processes (DPPs) are specific probability distributions over clouds of points that are used as models and computational tools across physics, probability, statistics, and more recently machine learning. Sampling from DPPs is a challenge and therefore we present DPPy, a Python toolbox that gathers known exact and approximate sampling algorithms. The project is hosted on GitHub and equipped with an extensive documentation. This documentation takes the form of a short survey of DPPs and relates each mathematical property with DPPy objects.

Streaming kernel regression with provably adaptive mean, variance, and regularization, [14]

We consider the problem of streaming kernel regression, when the observations arrive sequentially and the goal is to recover the underlying mean function, assumed to belong to an RKHS. The variance of the noise is not assumed to be known. In this context, we tackle the problem of tuning the regularization parameter adaptively at each time step, while maintaining tight confidence bounds estimates on the value of the mean function at each point. To this end, we first generalize existing results for finite-dimensional linear regression with fixed regularization and known variance to the kernel setup with a regularization parameter allowed to be a measurable function of past observations. Then, using appropriate self-normalized inequalities we build upper and lower bound estimates for the variance, leading to Bersntein-like concentration bounds. The later is used in

order to define the adaptive regularization. The bounds resulting from our technique are valid uniformly over all observation points and all time steps, and are compared against the literature with numerical experiments. Finally, the potential of these tools is illustrated by an application to kernelized bandits, where we revisit the Kernel UCB and Kernel Thompson Sampling procedures, and show the benefits of the novel adaptive kernel tuning strategy.

7.2. Applications

7.2.1. Dialogue Systems and Natural Language

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

End-to-End Automatic Speech Translation of Audiobooks, [24]

We investigate end-to-end speech-to-text translation on a corpus of audiobooks specifically augmented for this task. Previous works investigated the extreme case where source language transcription is not available during learning nor decoding , but we also study a midway case where source language transcription is available at training time only. In this case, a single model is trained to decode source speech into target text in a single pass. Experimental results show that it is possible to train compact and efficient end-to-end speech translation models in this setup. We also distribute the corpus and hope that our speech translation baseline on this corpus will be challenged in the future.

Visual Reasoning with Multi-hop Feature Modulation, [42]

Recent breakthroughs in computer vision and natural language processing have spurred interest in challenging multi-modal tasks such as visual question-answering and visual dialogue. For such tasks, one successful approach is to condition image-based convolutional network computation on language via Feature-wise Linear Modulation (FiLM) layers, i.e., per-channel scaling and shifting. We propose to generate the parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once, as in prior work. By alternating between attending to the language input and generating FiLM layer parameters, this approach is better able to scale to settings with longer input sequences such as dialogue. We demonstrate that multi-hop FiLM generation achieves state-of-the-art for the short input sequence task ReferIt-on-par with single-hop FiLM generation-while also significantly outperforming prior state-of-the-art and single-hop FiLM generation on the GuessWhat?! visual dialogue task.

7.2.2. Recommendation systems

Recurrent Neural Networks for Long and Short-Term Sequential Recommendation, [54]

Recommender systems objectives can be broadly characterized as modeling user preferences over short- or long-term time horizon. A large body of previous research studied long-term recommendation through dimensionality reduction techniques applied to the historical user-item interactions. A recently introduced session-based recommendation setting highlighted the importance of modeling short-term user preferences. In this task, Recurrent Neural Networks (RNN) have shown to be successful at capturing the nuances of user's interactions within a short time window. In this paper, we evaluate RNN-based models on both short-term and long-term recommendation tasks. Our experimental results suggest that RNNs are capable of predicting immediate as well as distant user interactions. We also find the best performing configuration to be a stacked RNN with layer normalization and tied item embeddings.

Fighting Boredom in Recommender Systems with Linear Reinforcement Learning, [43]

A common assumption in recommender systems (RS) is the existence of a best fixed recommendation strategy. Such strategy may be simple and work at the item level (e.g., in multi-armed bandit it is assumed one best fixed arm/item exists) or implement more sophisticated RS (e.g., the objective of A/B testing is to find the best fixed RS and execute it thereafter). We argue that this assumption is rarely verified in practice, as the recommendation process itself may impact the user's preferences. For instance, a user may get bored by a strategy, while she may gain interest again, if enough time passed since the last time that strategy was used. In this case, a better approach consists in alternating different solutions at the right frequency to fully exploit their potential. In this paper, we first cast the problem as a Markov decision process, where the rewards are a linear function of the recent history of actions, and we show that a policy considering the long-term influence of the recommendations may outperform both fixed-action and contextual greedy policies. We then introduce an extension of the UCRL algorithm (LINUCRL) to effectively balance exploration and exploitation in an unknown environment, and we derive a regret bound that is independent of the number of states. Finally, we empirically validate the model assumptions and the algorithm in a number of realistic scenarios.

7.2.3. Autonomous car

A Survey of State-Action Representations for Autonomous Driving, [51]

Approximate Robust Control of Uncertain Dynamical Systems, [40]

This work studies the design of safe control policies for large-scale non-linear systems operating in uncertain environments. In such a case, the robust control framework is a principled approach to safety that aims to maximize the worst-case performance of a system. However, the resulting optimization problem is generally intractable for non-linear systems with continuous states. To overcome this issue, we introduce two tractable methods that are based either on sampling or on a conservative approximation of the robust objective. The proposed approaches are applied to the problem of autonomous driving.

7.2.4. Software development

Correctness Attraction: A Study of Stability of Software Behavior Under Runtime Perturbation, [12]

Can the execution of a software be perturbed without breaking the correctness of the output? In this paper, we devise a novel protocol to answer this rarely investigated question. In an experimental study, we observe that many perturbations do not break the correctness in ten subject programs. We call this phenomenon "correctness attraction". The uniqueness of this protocol is that it considers a systematic exploration of the perturbation space as well as perfect oracles to determine the correctness of the output. To this extent, our findings on the stability of software under execution perturbations have a level of validity that has never been reported before in the scarce related work. A qualitative manual analysis enables us to set up the first taxonomy ever of the reasons behind correctness attraction.

This paper has attracted a significant interest in the SE community. This work has been invited for an oral presentation (along a 1 page summary) at the 40th International Conference on Software Engineering, the main conference in software engineering. It has then been invited on the IEEE Software review blog.

SMPyBandits: an Experimental Framework for Single and Multi-Players Multi-Arms Bandits Algorithms in Python, [46]

SMPyBandits is a package for numerical simulations on single-player and multi-players Multi-Armed Bandits (MAB) algorithms, written in Python (2 or 3). This library is the most complete open-source implementation of state-of-the-art algorithms tackling various kinds of sequential learning problems referred to as Multi-Armed Bandits. It is extensive, simple to use and maintain, with a clean and well documented codebase. It allows fast prototyping of experiments, with an easy configuration system and command-line options to customize experiments.

Lilian Besson developed a library for multi-armed bandit algorithms in Python for single and multi-player bandits.

7.2.5. Deep Learning

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

Feature-wise transformations, [13]

i-RevNet: Deep Invertible Networks, [32]

It is widely believed that the success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand. This is supported empirically by the difficulty of recovering images from their hidden representations, in most commonly used network architectures. In this paper we show via a one-to-one mapping that this loss of information is not a necessary condition to learn representations that generalize well on complicated problems, such as ImageNet. Via a cascade of homeomorphic layers, we build the i-RevNet, a network that can be fully inverted up to the final projection onto the classes, i.e. no information is discarded. Building an invertible architecture is difficult, for one, because the local inversion is ill-conditioned, we overcome this by providing an explicit inverse. An analysis of i-RevNets learned representations suggests an alternative explanation for the success of deep networks by a progressive contraction and linear separation with depth. To shed light on the nature of the model learned by the i-RevNet we reconstruct linear interpolations between natural image representations.

Compressing the Input for CNNs with the First-Order Scattering Transform, [33]

We study the first-order scattering transform as a candidate for reducing the signal processed by a convolutional neural network (CNN). We study this transformation and show theoretical and empirical evidence that in the case of natural images and sufficiently small translation invariance, this transform preserves most of the signal information needed for classification while substantially reducing the spatial resolution and total signal size. We show that cascading a CNN with this representation performs on par with ImageNet classification models commonly used in downstream tasks such as the ResNet-50. We subsequently apply our trained hybrid ImageNet model as a base model on a detection system, which has typically larger image inputs. On Pascal VOC and COCO detection tasks we deliver substantial improvements in the inference speed and training memory consumption compared to models trained directly on the input image.

Visual Reasoning with Multi-hop Feature Modulation, [42]

Recent breakthroughs in computer vision and natural language processing have spurred interest in challenging multi-modal tasks such as visual question-answering and visual dialogue. For such tasks, one successful approach is to condition image-based convolutional network computation on language via Feature-wise Linear Modulation (FiLM) layers, i.e., per-channel scaling and shifting. We propose to generate the parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once, as in prior work. By alternating between attending to the language input and generating FiLM layer parameters, this approach is better able to scale to settings with longer input sequences such as dialogue. We demonstrate that multi-hop FiLM generation achieves state-of-the-art for the short input sequence task ReferIt-on-par with single-hop FiLM generation-while also significantly outperforming prior state-of-the-art and single-hop FiLM generation on the GuessWhat?! visual dialogue task.

SPIRALS Project-Team

7. New Results

7.1. Software Product Lines for Setup and Adaptation of Multi-Cloud Computing Systems

In 2018, in the domain of cloud computing, we proposed a new software product line-based approach for managing the variability in order to automate the setup and adaptation of multi-cloud environments. Building such systems is still very challenging and time consuming due to the heterogeneity across cloud providers' offerings and the high-variability in the configuration of cloud providers. This variability is expressed by the large number of available services and the many different ways in which they can be combined and configured. In order to ensure correct setup of a multi-cloud environment, developers must be aware of service offerings and configuration options from multiple cloud providers. Our results enable to automatically generate a configuration or reconfiguration plan for a multi-cloud environment from a description of its requirements. The conducted experiments aim to assess the impact of the approach on the automated analysis of feature models and the feasibility of the approach to automate the setup and adaptation of multi-cloud environments. These results have been obtained in the context of the PhD thesis of Gustavo Sousa [12] defended in June 2018.

7.2. Automated Software Repair with Patch Generation in Production

In 2018, in the domain of automated software repair, we proposed new patch generation techniques. Patch creation is one of the most important actions in the life cycle of an application. Creating patches is a time-consuming task. Not only because it is difficult to create a sound and valid patch, but also because it requires the intervention of humans. Our work proposes new patch generation techniques that remove the human intervention. Our idea is to put as close as possible the patch generation in the production environment. We adopt this approach because the production environment contains all the data and human interactions that lead to the bug. We show how to exploit this data to detect bugs, generate and validate patches. We evaluate this approach on seven different benchmarks of real bugs collected from open-source projects. During the evaluation, we are particularly attentive to the number of generated patches, to their correctness, readability and to the time required for generating them. Our evaluation shows the applicability and feasibility of our approach to generate patches in the production environment without the intervention of a developer. These results have been obtained in the context of the PhD thesis of Thomas Durieux [11] defended in September 2018.

7.3. Flexible Framework for Elasticity in Cloud Computing

In 2018, in the domain of cloud computing, we proposed a new framework for managing elasticity. The main factor motivating the use of cloud is its ability to provide resources according to the customer needs or what is referred to as elasticity. Adapting cloud applications during their execution according to demand variation is nevertheless a challenging task. In addition, cloud elasticity is diverse and heterogeneous because it encompasses different approaches, policies, purposes, etc. In this work, three contributions are proposed: (1) an up-to-date state-of-the-art of the cloud elasticity for both virtual machines and containers, (2) ELASTIC-DOCKER, an approach to manage container elasticity including vertical elasticity, live migration, and elasticity combination between different virtualization techniques, and (3) MODEMO, a new unified standard-based, model-driven, highly extensible and reconfigurable framework that supports multiple elasticity policies, vertical and horizontal elasticity, different virtualization techniques and multiple cloud providers. These results have been obtained in the context of the PhD thesis of Yahya Al-Dhuraibi defended in December 2018.

7.4. Semantic Interoperability in Multi-Cloud Computing Systems

In 2018, in the domain of cloud computing, we proposed two major results related to semantic interoperability. First, an approach based on reverse-engineering to extract knowledge from the ambiguous textual documentation of cloud APIs and to enhance its representation using MDE techniques has been proposed. This approach is applied to Google Cloud Platform (GCP), where we provide GCP Model, a precise model-driven specification for GCP. GCP Model is automatically inferred from GCP textual documentation, conforms to the OCCIware Metamodel and is implemented within OCCIware Studio. It allows one to perform qualitative and quantitative analysis of the GCP documentation. Second, we have proposed the FCLOUDS framework to achieve semantic interoperability in multi-clouds, i.e., to identify the common concepts between cloud APIs and to reason over them. The FCLOUDS language is a formalization of OCCI concepts and operational semantics in Alloy formal specification language. To demonstrate the effectiveness of the FCLOUDS language, we formally specify thirteen case studies and verify their properties. Then, thanks to formal transformation rules and equivalence properties, we draw a precise alignment between my case studies, which promotes semantic interoperability in multi-clouds. These results have been obtained in the context of the PhD thesis of Stéphanie Challita defended in December 2018.