Activity Report 2018

Project-Team DATASHAPE

Understanding the shape of data

RESEARCH CENTERS
Saclay - Île-de-France
Sophia Antipolis - Méditerranée

THEME
Algorithmics, Computer Algebra and Cryptology

Table of contents

Project-Team DATASHAPE

Creation of the Team: 2016 January 01, updated into Project-Team: 2016 January 01

Keywords:

Computer Science and Digital Science:

A3. - Data and knowledge
A3.4. - Machine learning and statistics
A7.1. - Algorithms
A8. - Mathematics of computing
A8.1. - Discrete mathematics, combinatorics
A8.3. - Geometry, Topology
A9. - Artificial intelligence

Other Research Topics and Application Domains:

B1. - Life sciences
B2. - Health
B5. - Industry of the future
B9. - Society and Knowledge
B9.5. - Sciences

1. Team, Visitors, External Collaborators

Research Scientists

Jean-Daniel Boissonnat [Inria, Senior Researcher, Sophia Antipolis - Méditerranée, HDR]
Frédéric Chazal [Team leader, Inria, Senior Researcher, Saclay - Île-de-France, HDR]
David Cohen-Steiner [Inria, Researcher, Sophia Antipolis - Méditerranée]
Kunal Dutta [Inria, Starting Research Position, Sophia Antipolis - Méditerranée]
Marc Glisse [Inria, Researcher, Saclay - Île-de-France]
Clément Maria [Inria, Researcher, from Feb 2018, Sophia Antipolis - Méditerranée]
Steve Oudot [Inria, Researcher, Saclay - Île-de-France, HDR]
Mathijs Wintraecken [Inria, Starting Research Position, Sophia Antipolis - Méditerranée]
Miroslav Kramár [Inria, Advanced Research Position, Saclay - Île-de-France]

Faculty Members

Guilherme Dias Da Fonseca [Univ d'Auvergne, Associate Professor, until Oct 2018, Sophia Antipolis - Méditerranée]
Clément Levrard [Univ Denis Diderot, Associate Professor, from Sep 2018, Saclay - Île-de-France]

Post-Doctoral Fellows

Mathieu Carrière [Inria, until Jul 2018, Saclay - Île-de-France]
Hariprasad Kannan [Inria, from Oct 2018, Saclay - Île-de-France]
Jisu Kim [Inria, from Nov 2018, Saclay - Île-de-France]
Martin Royer [Inria, from Dec 2018, Saclay - Île-de-France]

PhD Students

Bertrand Beaufils [Sysnav, industrial grant, Saclay - Île-de-France]
Nicolas Berkouk [Inria, Saclay - Île-de-France]
Claire Brécheteau [Univ Paris-Sud, until Aug 2018, Saclay - Île-de-France]
Jeremy Sylvain Cochoy [Univ Paris-Sud, until Aug 2018, Saclay - Île-de-France]

Vincent Divol [Univ Paris-Sud, from Sep 2018, Saclay - Île-de-France]
Francois Godi [Univ de Nice - Sophia Antipolis, until Sep 2018, Sophia Antipolis - Méditerranée]
Siargey Kachanovich [Inria, Sophia Antipolis - Méditerranée]
Hariprasad Kannan [Inria, from Jul 2018 until Sep 2018, Saclay - Île-de-France]
Théo Lacombe [Ecole polytechnique, Saclay - Île-de-France]
Siddharth Pritam [Inria, Sophia Antipolis - Méditerranée]
Owen Rouille [Inria, from Sep 2018, Sophia Antipolis - Méditerranée]
Martin Royer [Inria, from Oct 2018 until Nov 2018, Saclay - Île-de-France]
Raphaël Tinarrage [Ecole Normale Supérieure Cachan, Saclay - Île-de-France]

Technical staff

Thomas Bonis [Inria, from Jul 2018, Saclay - Île-de-France]
Vincent Rouvreau [Inria, Saclay - Île-de-France]

Interns

Shreya Arya [Inria, from Jul 2018, Sophia Antipolis - Méditerranée]
Vincent Divol [Ecole Normale Supérieure Cachan, until Feb 2018, Saclay - Île-de-France]
Etienne Lasalle [Ecole Normale Supérieure Cachan, from Apr 2018 until Jul 2018, Saclay - Île-de-France]
Louis Pujol [Inria, from May 2018 until Jul 2018, Saclay - Île-de-France]
Owen Rouille [Ecole Normale Supérieure Rennes, from Feb 2018 until Jun 2018, Sophia Antipolis - Méditerranée]
Younes Zine [Univ de Rennes I, from Apr 2018 until Sep 2018, Saclay - Île-de-France]
Shourya Pandey [IIT Bombay, from May 2018 until Jul 2018, Sophia Antipolis - Méditerranée]

Administrative Assistants

Stephanie Aubin [Inria, until Sep 2018, Saclay - Île-de-France]
Florence Barbara [Inria, Sophia Antipolis - Méditerranée]
Sophie Honnorat [Inria, from Sep 2018, Sophia Antipolis - Méditerranée]
Hélèna Kutniak [Inria, from Oct 2018, Saclay - Île-de-France]

Visiting Scientists

Vincent Divol [Ecole Normale Supérieure Cachan, from Mar 2018 until Aug 2018, Saclay - Île-de-France]
Arijit Ghosh [Indian Statistical Institute, Kolkata, India, from Apr until May 2018, Dec 2018, Sophia Antipolis - Méditerranée]
Ilaria Giulini [Univ Pierre et Marie Curie, until Sep 2018, Saclay - Île-de-France]
Wolfgang Polonik [UC Davis, from Sep 2018 until Oct 2018, Saclay - Île-de-France]
Hannah Santa Cruz Baur [CIMAT, from Jun 2018 until Aug 2018, Sophia Antipolis - Méditerranée]
Hannah Schreiber [TU Graz, until Jul 2018, Sophia Antipolis - Méditerranée]
Yitchzak Solomon [Brown University, from Sep 2018, Saclay - Île-de-France]
Ramsay Dyer [Berkeley Publishing, from Apr until May 2018, Dec 2018, Sophia Antipolis - Méditerranée]
Mariette Yvinec [Retraitée, Sophia Antipolis - Méditerranée, HDR]

External Collaborators

Maksims Ovsjanikovs [Ecole polytechnique, Professor, Saclay - Île-de-France]
Bertrand Michel [Ecole Centrale de Nantes, Saclay - Île-de-France]

2. Overall Objectives

2.1. Overall Objectives

DataShape is a research project in Topological Data Analysis (TDA), a recent field whose aim is to uncover, understand and exploit the topological and geometric structure underlying complex and possibly high dimensional data. The DATASHAPE project gathers a unique variety of expertise that allows it to embrace the mathematical, statistical, algorithmic and applied aspects of the field in a common framework ranging from fundamental theoretical studies to experimental research and software development.

The expected output of DATASHAPE is two-fold. First, we intend to set-up and develop the mathematical, statistical and algorithmic foundations of Topological and Geometric Data Analysis. Second, we intend to develop the Gudhi platform in order to provide an efficient state-of-the-art toolbox for the understanding of the topology and geometry of data.

3. Research Program

3.1. Algorithmic aspects of topological and geometric data analysis

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

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

3.2. Statistical aspects of topological and geometric data analysis

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

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

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

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

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

3.3. Topological approach for multimodal data processing

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

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

3.4. Experimental research and software development

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

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

Some of the algorithms implemented on the platform will also be interfaced to other software platform, such as the R software [1] for statistical computing, and languages such as Python in order to make them usable in combination with other data analysis and machine learning tools. A first attempt in this direction has been done with the creation of an R package called TDA in collaboration with the group of Larry Wasserman at Carnegie Mellon University (Inria Associated team CATS) that already includes some functionalities of the GUDHI library and implements some joint results between our team and the CMU team. A similar interface with the Python language is also considered a priority. To go even further towards helping users, we will provide utilities that perform the most common tasks without requiring any programming at all.

4. Application Domains

4.1. Main application domains

Our work is mostly of a fundamental mathematical and algorithmic nature but finds a variety of applications in data analysis, e.g., in material science, biology, sensor networks, 3D shape analysis and processing, to name a few.

More specifically, DATASHAPE is working on the analysis of trajectories obtained from inertial sensors (PhD thesis of Bertrand Beaufils with Sysnav) and, more generally on the development of new TDA methods for Machine Learning and Artificial Intelligence for (multivariate) time-dependent data from various kinds of sensors in collaboration with Fujitsu.

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

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Books

- Jean-Daniel Boissonnat, Frédéric Chazal, Mariette Yvinec. *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics, vol. 57, Cambridge University Press, 2018.

5.1.2. Awards

- Mathieu Carrière was awarded the Prix de thèse solennel Thiessé de Rosemont / Schneider in Mathematics by the Chancellerie des Universités de Paris for his Ph.D. work under Steve Oudot's supervision (Ph.D. funded by ERC grant Gudhi), December 2018.

6. New Software and Platforms

6.1. GUDHI

Geometric Understanding in Higher Dimensions
KEYWORDS: Computational geometry - Topology
SCIENTIFIC DESCRIPTION: The current release of the GUDHI library includes: – Data structures to represent, construct and manipulate simplicial and cubical complexes. – Algorithms to compute simplicial complexes from point cloud data. – Algorithms to compute persistent homology and multi-field persistent homology. – Simplification methods via implicit representations.
FUNCTIONAL DESCRIPTION: The GUDHI open source library will provide the central data structures and algorithms that underly applications in geometry understanding in higher dimensions. It is intended to both help the development of new algorithmic solutions inside and outside the project, and to facilitate the transfer of results in applied fields.
NEWS OF THE YEAR: - Cover complex - Representation of persistence diagrams - Cech complex - weighted periodic 3d alpha-complex - sparse Rips complex - debian / docker / conda-forge packages

- Participants: Clément Maria, François Godi, David Salinas, Jean-Daniel Boissonnat, Marc Glisse, Mariette Yvinec, Pawel Dlotko, Siargey Kachanovich and Vincent Rouvreau
- Contact: Jean-Daniel Boissonnat
- URL: http://gudhi.gforge.inria.fr/

7. New Results

7.1. Algorithmic aspects of topological and geometric data analysis

7.1.1. DTM-based filtrations

Participants: Frédéric Chazal, Marc Glisse, Raphaël Tinarrage.

In collaboration with H. Anai, Y. Ike, H. Inakoshi and Y. Umeda of Fujitsu.

Despite strong stability properties, the persistent homology of filtrations classically used in Topological Data Analysis, such as, e.g. the Čech or Vietoris-Rips filtrations, are very sensitive to the presence of outliers in the data from which they are computed. In this paper [33], we introduce and study a new family of filtrations, the DTM-filtrations, built on top of point clouds in the Euclidean space which are more robust to noise and outliers. The approach adopted in this work relies on the notion of distance-to-measure functions, and extends some previous work on the approximation of such functions.

7.1.2. Persistent Homology with Dimensionality Reduction: k-Distance vs Gaussian Kernels

Participants: Shreya Arya, Jean-Daniel Boissonnat, Kunal Dutta.

We investigate the effectiveness of dimensionality reduction for computing the persistent homology for both k-distance and kernel distance [34]. For k-distance, we show that the standard Johnson-Lindenstrauss reduction preserves the k-distance, which preserves the persistent homology upto a $(1-\varepsilon)^{-1}$ factor with target dimension $O(k \log n/\varepsilon^2)$. We also prove a concentration inequality for sums of dependent chi-squared random variables, which, under some conditions, allows the persistent homology to be preserved in $O(\log n/\varepsilon^2)$ dimensions. This answers an open question of Sheehy. For Gaussian kernels, we show that the standard Johnson-Lindenstrauss reduction preserves the persistent homology up to an $4(1-\epsilon)^{-1}$ factor.

7.1.3. Computing Persistent Homology of Flag Complexes via Strong Collapses

Participants: Jean-Daniel Boissonnat, Siddharth Pritam.

In collaboration with Divyansh Pareek (Indian Institute of Technology Bombay, India)

We introduce a fast and memory efficient approach to compute the persistent homology (PH) of a sequence of simplicial complexes. The basic idea is to simplify the complexes of the input sequence by using strong collapses, as introduced by J. Barmak and E. Miniam [DCG (2012)], and to compute the PH of an induced sequence of reduced simplicial complexes that has the same PH as the initial one. Our approach has several salient features that distinguishes it from previous work. It is not limited to filtrations (i.e. sequences of nested simplicial subcomplexes) but works for other types of sequences like towers and zigzags. To strong collapse a simplicial complex, we only need to store the maximal simplices of the complex, not the full set of all its simplices, which saves a lot of space and time. Moreover, the complexes in the sequence can be strong collapsed independently and in parallel. Finally, we can compromize between precision and time by choosing the number of simplicial complexes of the sequence we strong collapse. As a result and as demonstrated by numerous experiments on publicly available data sets, our approach is extremely fast and memory efficient in practice [27].

7.1.4. Strong Collapse for Persistence

Participants: Jean-Daniel Boissonnat, Siddharth Pritam.

In this paper, we build on the initial success of and show that further decisive progress can be obtained if one restricts the family of simplicial complexes to flag complexes. Flag complexes are fully characterized by their graph (or 1-skeleton), the other faces being obtained by computing the cliques of the graph. Hence, a flag complex can be represented by its graph, which is a very compact representation. Flag complexes are very popular and, in particular, Vietoris-Rips complexes are by far the most widely simplicial complexes used in Topological Data Analysis. It has been shown in that the persistent homology of Vietoris-Rips filtrations can be computed very efficiently using strong collapses. However, most of the time was devoted to computing the maximal cliques of the complex prior to their strong collapse. In this paper [37], we observe that the reduced complex obtained by strong collapsing a flag complex is itself a flag complex. Moreover, this reduced complex can be computed using only the 1-skeleton (or graph) of the complex, not the set of its maximal cliques. Finally, we show how to compute the equivalent filtration of the sequence of reduced flag simplicial complexes using again only 1-skeletons. x On the theory side, we show that strong collapses of flag complexes can be computed in time $O(v^2k^2)$ where v is the number of vertices of the complex and k the maximal degree of its graph. The algorithm described in this paper has been implemented and the code will be soon released in the Gudhi library. Numerous experiments show that our method outperforms previous methods, e.g. Ripser.

7.1.5. Triangulating submanifolds: An elementary and quantified version of Whitney's method

Participants: Jean-Daniel Boissonnat, Siargey Kachanovich, Mathijs Wintraecken.

We quantize Whitney's construction to prove the existence of a triangulation for any C^2 manifold, so that we get an algorithm with explicit bounds. We also give a new elementary proof, which is completely geometric [36].

7.1.6. Randomized incremental construction of Delaunay triangulations of nice point sets

Participants: Jean-Daniel Boissonnat, Kunal Dutta, Marc Glisse.

In collaboration with Olivier Devillers (Inria, CNRS, Loria, Université de Lorraine).

Randomized incremental construction (RIC) is one of the most important paradigms for building geometric data structures. Clarkson and Shor developed a general theory that led to numerous algorithms that are both simple and efficient in theory and in practice.

Randomized incremental constructions are most of the time space and time optimal in the worst-case, as exemplified by the construction of convex hulls, Delaunay triangulations and arrangements of line segments.

However, the worst-case scenario occurs rarely in practice and we would like to understand how RIC behaves when the input is nice in the sense that the associated output is significantly smaller than in the worst-case. For example, it is known that the Delaunay triangulations of nicely distributed points in $\mathbb{R}^d$ or on polyhedral surfaces in $\mathbb{R}^3$ has linear complexity, as opposed to a worst-case complexity of $\Theta(n^{\lfloor d/2 \rfloor})$ in the first case and quadratic in the second. The standard analysis does not provide accurate bounds on the complexity of such cases and we aim at establishing such bounds in this paper [35]. More precisely, we will show that, in the two cases above and variants of them, the complexity of the usual RIC is $O(n \log n)$, which is optimal. In other words, without any modification, RIC nicely adapts to good cases of practical value.

Along the way, we prove a probabilistic lemma for sampling without replacement, which may be of independent interest.

7.1.7. Approximate Polytope Membership Queries

Participant: Guilherme Da Fonseca.

In collaboration with Sunil Arya (Hong Kong University of Science and Technology) and David Mount (University of Maryland).

In the polytope membership problem, a convex polytope K in $\mathbb{R}^d$ is given, and the objective is to preprocess K into a data structure so that, given any query point $q \in \mathbb{R}^d$, it is possible to determine efficiently whether $q \in K$. We consider this problem in an approximate setting. Given an approximation parameter ϵ, the query can be answered either way if the distance from q to K's boundary is at most ϵ times K's diameter. We assume that the dimension d is fixed, and K is presented as the intersection of n halfspaces. Previous solutions to approximate polytope membership were based on straightforward applications of classic polytope approximation techniques by Dudley (1974) and Bentley et al. (1982). The former is optimal in the worst-case with respect to space, and the latter is optimal with respect to query time. We present four main results. First, we show how to combine the two above techniques to obtain a simple space-time trade-off. Second, we present an algorithm that dramatically improves this trade-off. In particular, for any constant $\alpha \geq 4$, this data structure achieves query time roughly $O(1/\epsilon^{(d-1)/\alpha})$ and space roughly $O(1/\epsilon^{(d-1)(1-\Omega(\log \alpha))/\alpha})$. We do not know whether this space bound is tight, but our third result shows that there is a convex body such that our algorithm achieves a space of at least $\Omega(1/\epsilon^{(d-1)(1-O(\sqrt{\alpha}))/\alpha})$. Our fourth result shows that it is possible to reduce approximate Euclidean nearest neighbor searching to approximate polytope membership queries. Combined with the above results, this provides significant improvements to the best known space-time trade-offs for approximate nearest neighbor searching in $\mathbb{R}^d$. For example, we show that it is possible to achieve a query time of roughly $O(\log n + 1/\epsilon^{d/4})$ with space roughly $O(n/\epsilon^{d/4})$, thus reducing by half the exponent in the space bound [11].

7.1.8. Approximate Convex Intersection Detection with Applications to Width and Minkowski Sums

Participant: Guilherme Da Fonseca.

In collaboration with Sunil Arya (Hong Kong University of Science and Technology) and David Mount (University of Maryland).

Approximation problems involving a single convex body in d-dimensional space have received a great deal of attention in the computational geometry community. In contrast, works involving multiple convex bodies are generally limited to dimensions $d \leq 3$ and/or do not consider approximation. In this paper, we consider approximations to two natural problems involving multiple convex bodies: detecting whether two polytopes intersect and computing their Minkowski sum. Given an approximation parameter $\epsilon > 0$, we show how to independently preprocess two polytopes A, B into data structures of size $O(1/\epsilon^{(d-1)/2})$ such that we can answer in polylogarithmic time whether A and B intersect approximately. More generally, we can answer this for the images of A and B under affine transformations. Next, we show how to ϵ-approximate the Minkowski sum of two given polytopes defined as the intersection of n halfspaces in $O(n \log(1/\epsilon) + 1/\epsilon^{(d-1)/2+\alpha})$ time, for any constant $\alpha > 0$. Finally, we present a surprising impact of these results to a well studied problem that considers a single convex body. We show how to ϵ-approximate the width of a set of n points in $O(n \log(1/\epsilon) + 1/\epsilon^{(d-1)/2+\alpha})$ time, for any constant $\alpha > 0$, a major improvement over the previous bound of roughly $O(n + 1/\epsilon^{d-1})$ time [22].

7.1.9. Approximating the Spectrum of a Graph

Participant: David Cohen-Steiner.

In collaboration with Weihao Kong (Stanford University), Christian Sohler (TU Dortmund) and Gregory Valiant (Stanford University).

The spectrum of a network or graph $G = (V, E)$ with adjacency matrix A , consists of the eigenvalues of the normalized Laplacian $L = I - D^{-1/2}AD^{-1/2}$. This set of eigenvalues encapsulates many aspects of the structure of the graph, including the extent to which the graph posses community structures at multiple scales. We study the problem of approximating the spectrum, $\lambda = (\lambda_1, \cdots, \lambda_{|V|})$, of G in the regime where the graph is too large to explicitly calculate the spectrum. We present a sublinear time algorithm that, given the ability to query a random node in the graph and select a random neighbor of a given node, computes a succinct representation of an approximation $\widetilde{\lambda} = (\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_{|V|})$, such that $\|\widetilde{\lambda} - \lambda\|_1 \leq \varepsilon |V|$. Our algorithm has query complexity and running time $\exp(O(1/\varepsilon))$, which is independent of the size of the graph, $|V|$. We demonstrate the practical viability of our algorithm on synthetically generated graphs, and on 15 different real-world graphs from the Stanford Large Network Dataset Collection, including social networks, academic collaboration graphs, and road networks. For the smallest of these graphs, we are able to validate the accuracy of our algorithm by explicitly calculating the true spectrum; for the larger graphs, such a calculation is computationally prohibitive. The spectra of these real-world networks reveal insights into the structural similarities and differences between them, illustrating the potential value of our algorithm for efficiently approximating the spectrum of large networks [29].

7.1.10. Spectral Properties of Radial Kernels and Clustering in High Dimensions

Participants: David Cohen-Steiner, Alba Chiara de Vitis.

In this paper [40], we study the spectrum and the eigenvectors of radial kernels for mixtures of distributions in $\mathbb{R}^n$. Our approach focuses on high dimensions and relies solely on the concentration properties of the components in the mixture. We give several results describing of the structure of kernel matrices for a sample drawn from such a mixture. Based on these results, we analyze the ability of kernel PCA to cluster high dimensional mixtures. In particular, we exhibit a specific kernel leading to a simple spectral algorithm for clustering mixtures with possibly common means but different covariance matrices. This algorithm will succeed if the angle between any two covariance matrices in the mixture (seen as vectors in $\mathbb{R}^{n^2}$) is larger than $\Omega(n^{-1/6} \log^{5/3} n)$. In particular, the required angular separation tends to 0 as the dimension tends to infinity. To the best of our knowledge, this is the first polynomial time algorithm for clustering such mixtures beyond the Gaussian case.

7.1.11. Exact computation of the matching distance on 2-parameter persistence modules

Participant: Steve Oudot.

In collaboration with Michael Kerber (T.U. Graz) and Michael Lesnick (SUNY).

The matching distance is a pseudometric on multi-parameter persistence modules, defined in terms of the weighted bottleneck distance on the restriction of the modules to affine lines. It is known that this distance is stable in a reasonable sense, and can be efficiently approximated, which makes it a promising tool for practical applications. In [44] we show that in the 2-parameter setting, the matching distance can be computed exactly in polynomial time. Our approach subdivides the space of affine lines into regions, via a line arrangement. In each region, the matching distance restricts to a simple analytic function, whose maximum is easily computed. As a byproduct, our analysis establishes that the matching distance is a rational number, if the bigrades of the input modules are rational.

7.1.12. A Comparison Framework for Interleaved Persistence Modules

Participant: Miroslav Kramár.

In collaboration with Rachel Levanger (UPenn), Shaun Harker and Konstantin Mischaikow (Rutgers).

In [43], we present a generalization of the induced matching theorem of [1] and use it to prove a generalization of the algebraic stability theorem for R-indexed pointwise finite-dimensional persistence modules. Via numerous examples, we show how the generalized algebraic stability theorem enables the computation of rigorous error bounds in the space of persistence diagrams that go beyond the typical formulation in terms of bottleneck (or log bottleneck) distance.

7.1.13. Discrete Morse Theory for Computing Zigzag Persistence

Participant: Clément Maria.

In collaboration with Hannah Schreiber (Graz University of Technology, Austria)

We introduce a framework to simplify zigzag filtrations of general complexes using discrete Morse theory, in order to accelerate the computation of zigzag persistence. Zigzag persistence is a powerful algebraic generalization of persistent homology. However, its computation is much slower in practice, and the usual optimization techniques cannot be used to compute it. Our approach is different in that it preprocesses the filtration before computation. Using discrete Morse theory, we get a much smaller zigzag filtration with same persistence. The new filtration contains general complexes. We introduce new update procedures to modify on the fly the algebraic data (the zigzag persistence matrix) under the new combinatorial changes induced by the Morse reduction. Our approach is significantly faster in practice [45].

7.2. Statistical aspects of topological and geometric data analysis

7.2.1. Robust Bregman Clustering

Participants: Claire Brécheteau, Clément Levrard.

In collaboration with Aurélie Fischer (Université Paris-Diderot).

Using a trimming approach, in [38], we investigate a k-means type method based on Bregman divergences for clustering data possibly corrupted with clutter noise. The main interest of Bregman divergences is that the standard Lloyd algorithm adapts to these distortion measures, and they are well-suited for clustering data sampled according to mixture models from exponential families. We prove that there exists an optimal codebook, and that an empirically optimal codebook converges a.s. to an optimal codebook in the distortion sense. Moreover, we obtain the sub-Gaussian rate of convergence for k-means $1 \sqrt{n}$ under mild tail assumptions. Also, we derive a Lloyd-type algorithm with a trimming parameter that can be selected from data according to some heuristic, and present some experimental results.

7.2.2. Statistical analysis and parameter selection for Mapper

Participants: Mathieu Carrière, Bertrand Michel, Steve Oudot.

In [15] we study the question of the statistical convergence of the 1-dimensional Mapper to its continuous analogue, the Reeb graph. We show that the Mapper is an optimal estimator of the Reeb graph, which gives, as a byproduct, a method to automatically tune its parameters and compute confidence regions on its topological features, such as its loops and flares. This allows to circumvent the issue of testing a large grid of parameters and keeping the most stable ones in the brute-force setting, which is widely used in visualization, clustering and feature selection with the Mapper.

7.2.3. A Fuzzy Clustering Algorithm for the Mode-Seeking Framework

Participants: Thomas Bonis, Steve Oudot.

In [13] we propose a new soft clustering algorithm based on the mode-seeking framework. Given a point cloud in $\mathbb{R}^d$, we define regions of high density that we call cluster cores, then we implement a random walk on a neighborhood graph built on top of the data points. This random walk is designed in such a way that it is attracted by high-density regions, the intensity of the attraction being controlled by a temperature parameter $\beta > 0$. The membership of a point to a given cluster is then the probability for the random walk starting at this point to hit the corresponding cluster core before any other. While many properties of random walks (such as hitting times, commute distances, etc) are known to eventually encode purely local information when the number of data points grows to infinity, the regularization introduced by the use of cluster cores allows the output of our algorithm to converge to quantities involving the global structure of the underlying density function. Empirically, we show how the choice of β influences the behavior of our algorithm: for small values of β the result is really close to hard mode-seeking, while for values of β close to 1 the result is similar to the output of the (soft) spectral clustering. We also demonstrate the scalability of our approach experimentally.

7.2.4. Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport

Participants: Théo Lacombe, Steve Oudot.

In collaboration with Marco Cuturi (ENSAE).

Persistence diagrams (PDs) are at the core of topological data analysis. They provide succinct descriptors encoding the underlying topology of sophisticated data. PDs are backed-up by strong theoretical results regarding their stability and have been used in various learning contexts. However, they do not live in a space naturally endowed with a Hilbert structure where natural metrics are not even differentiable, thus not suited to optimization process. Therefore, basic statistical notions such as the barycenter of a finite sample of PDs are not properly defined. In [30] we provide a theoretically good and computationally tractable framework to estimate the barycenter of a set of persistence diagrams. This construction is based on the theory of Optimal Transport (OT) and endows the space of PDs with a metric inspired from regularized Wasserstein distances.

7.2.5. The k-PDTM : a coreset for robust geometric inference

Participants: Claire Brécheteau, Clément Levrard.

Analyzing the sub-level sets of the distance to a compact sub-manifold of $\mathbb{R}^d$ is a common method in TDA to understand its topology. The distance to measure (DTM) was introduced by Chazal, Cohen-Steiner and Mérigot to face the non-robustness of the distance to a compact set to noise and outliers. This function makes possible the inference of the topology of a compact subset of $\mathbb{R}^d$ from a noisy cloud of n points lying nearby in the Wasserstein sense. In practice, these sub-level sets may be computed using approximations of the DTM such as the q-witnessed distance or other power distance. These approaches lead eventually to compute the homology of unions of n growing balls, that might become intractable whenever n is large. To simultaneously face the two problems of large number of points and noise, we introduce in [39] the k-power distance to measure (k-PDTM). This new approximation of the distance to measure may be thought of as a k-coreset based approximation of the DTM. Its sublevel sets consist in union of k-balls, $k << n$, and this distance is also proved robust to noise. We assess the quality of this approximation for k possibly dramatically smaller than n, for instance $k = n13$ is proved to be optimal for 2-dimensional shapes. We also provide an algorithm to compute this k-PDTM.

7.2.6. The density of expected persistence diagrams and its kernel based estimation

Participants: Frédéric Chazal, Vincent Divol.

Persistence diagrams play a fundamental role in Topological Data Analysis where they are used as topological descriptors of filtrations built on top of data. They consist in discrete multisets of points in the plane $\mathbb{R}^2$ that can equivalently be seen as discrete measures in $\mathbb{R}^2$. When the data come as a random point cloud, these discrete measures become random measures whose expectation is studied in this paper. In [28] we first show that for a wide class of filtrations, including the Čech and Rips-Vietoris filtrations, the expected persistence diagram, that is a deterministic measure on $\mathbb{R}^2$, has a density with respect to the Lebesgue measure. Second, building on the previous result we show that the persistence surface recently introduced by Adams et al can be seen as a kernel estimator of this density. We propose a cross-validation scheme for selecting an optimal bandwidth, which is proven to be a consistent procedure to estimate the density.

7.2.7. On the choice of weight functions for linear representations of persistence diagrams

Participant: Vincent Divol.

In collaboration with Wolfgang Polonik (UC Davis)

Persistence diagrams are efficient descriptors of the topology of a point cloud. As they do not naturally belong to a Hilbert space, standard statistical methods cannot be directly applied to them. Instead, feature maps (or representations) are commonly used for the analysis. A large class of feature maps, which we call linear, depends on some weight functions, the choice of which is a critical issue. An important criterion to choose a weight function is to ensure stability of the feature maps with respect to Wasserstein distances on diagrams. In [42], we improve known results on the stability of such maps, and extend it to general weight functions. We also address the choice of the weight function by considering an asymptotic setting; assume that X_n is an i.i.d. sample from a density on $[0,1]^d$. For the Cech and Rips filtrations, we characterize the weight functions for which the corresponding feature maps converge as n approaches infinity, and by doing so, we prove laws of large numbers for the total persistence of such diagrams. Both approaches lead to the same simple heuristic for tuning weight functions: if the data lies near a d-dimensional manifold, then a sensible choice of weight function is the persistence to the power α with $\alpha \geq d$.

7.2.8. Estimating the Reach of a Manifold

Participants: Frédéric Chazal, Bertrand Michel.

In collaboration with E. Aamari (CNRS Paris 7), J.Kim, A. Rinaldo and L. Wasserman (Carnegie Mellon University).

Various problems in manifold estimation make use of a quantity called the reach, denoted by τ_M, which is a measure of the regularity of the manifold. [32] is the first investigation into the problem of how to estimate the reach. First, we study the geometry of the reach through an approximation perspective. We derive new geometric results on the reach for submanifolds without boundary. An estimator $\widehat{\tau}$ of τ_M is proposed in a framework where tangent spaces are known, and bounds assessing its efficiency are derived. In the case of i.i.d. random point cloud $\mathbb{X}_n$, $\tau(\mathbb{X}_n)$ is showed to achieve uniform expected loss bounds over a $\mathcal{C}^3$-like model. Finally, we obtain upper and lower bounds on the minimax rate for estimating the reach.

7.2.9. Robust Topological Inference: Distance To a Measure and Kernel Distance

Participants: Frédéric Chazal, Bertrand Michel.

In collaboration with B. Fasy (Univ. Montana) and F. Lecci, A. Rinaldo and L. Wasserman (Carnegie Mellon University).

Let P be a distribution with support S. The salient features of S can be quantified with persistent homology, which summarizes topological features of the sublevel sets of the distance function (the distance of any point x to S). Given a sample from P we can infer the persistent homology using an empirical version of the distance function. However, the empirical distance function is highly non-robust to noise and outliers. Even one outlier is deadly. The distance-to-a-measure (DTM), introduced by Chazal et al. (2011), and the kernel distance, introduced by Phillips et al. (2014), are smooth functions that provide useful topological information but are robust to noise and outliers. Chazal et al. (2015) derived concentration bounds for DTM. Building on these results, in [16], we derive limiting distributions and confidence sets, and we propose a method for choosing tuning parameters.

7.3. Topological approach for multimodal data processing

7.3.1. Barcode Embeddings for Metric Graphs

Participants: Steve Oudot, Yitchzak Solomon.

Stable topological invariants are a cornerstone of persistence theory and applied topology, but their discriminative properties are often poorly-understood. In [46] we study a rich homology-based invariant first defined by Dey, Shi, and Wang, which we think of as embedding a metric graph in the barcode space. We prove that this invariant is locally injective on the space of metric graphs and globally injective on a GH-dense subset. Moreover, we define a new topology on MGraphs, which we call the fibered topology, for which the barcode transform is injective on a generic (open and dense) subset.

7.3.2. Inverse Problems in Topological Persistence: a Survey

Participants: Steve Oudot, Yitchzak Solomon.

In [47] we review the literature on inverse problems in topological persistence theory. The first half of the survey is concerned with the question of surjectivity, i.e. the existence of right inverses, and the second half focuses on injectivity, i.e. left inverses. Throughout, we highlight the tools and theorems that underlie these advances, and direct the reader's attention to open problems, both theoretical and applied.

7.4. Experimental research and software development

7.4.1. Activity recognition from stride detection: a machine learning approach based on geometric patterns and trajectory reconstruction.

Participants: Bertrand Beaufils, Frédéric Chazal, Bertrand Michel.

In collaboration with M. Grelet (Sysnav).

In [23] algorithm for activity recognition is proposed using inertial sensors worn on the ankle. This innovative approach based on geometric patterns uses a stride detector that can detect both normal walking strides and atypical strides such as small steps, side steps and backward walking that existing methods struggle to detect. It is also robust in critical situations, when for example the wearer is sitting and moving the ankle, while most algorithms in the literature would wrongly detect strides. A technique inspired by Zero Velocity Update is used on the stride detection to compute the trajectory of the device. It allows to compute relevant features for the activity recognition learning task. Compared to most algorithms in the literature, this method does not use fixed-size sliding window that could be too short to provide enough information or too long and leads to overlapping issue when the window covers two different activities.

7.4.2. Dynamics of silo deformation under granular discharge

Participant: Miroslav Kramár.

In collaboration with Claudia Colonnello.

In [17], we use Topological Data Analysis to study the post buckling behavior of laboratory scale cylindrical silos under gravity driven granular discharges. Thin walled silos buckle during the discharge if the initial height of the granular column is large enough. The deformation of the silo is reversible as long as the filling height does not exceed a critical value, Lc. Beyond this threshold the deformation becomes permanent and the silo often collapses. We study the dynamics of reversible and irreversible deformation processes, varying the initial filling height around Lc. We find that all reversible processes exhibit striking similarities and they alternate between regimes of slow and fast dynamics. The patterns that occur at the beginning of irreversible deformation processes are topologically very similar to those that arise during reversible processes. However, the dynamics of reversible and irreversible processes is significantly different. In particular, the evolution of irreversible processes is much faster. This allows us to make an early prediction of the collapse of the silo based solely on observations of the deformation patterns.

7.4.3. Characterizing Granular Networks Using Topological Metrics

Participant: Miroslav Kramár.

In collaboration with Joshua Dijksman (Duke Physics), Lenka Kovalcinova and Lou Kondic (NJIT), Jie Ren (Merck Research Lab), Robert Behringer (Duke), and Konstantin Mischaikow (Rutgers).

In [18], we carry out a direct comparison of experimental and numerical realizations of the exact same granular system as it undergoes shear jamming. We adjust the numerical methods used to optimally represent the experimental settings and outcomes up to microscopic contact force dynamics. Measures presented here range form microscopic, through mesoscopic to system-wide characteristics of the system. Topological properties of the mesoscopic force networks provide a key link between mi-cro and macro scales. We report two main findings: the number of particles in the packing that have at least two contacts is a good predictor for the mechanical state of the system, regardless of strain history and packing density. All measures explored in both experiments and numerics, including stress tensor derived measures and contact numbers depend in a universal manner on the fraction of non-rattler particles, fNR. The force network topology also tends to show this universality, yet the shape of the master curve depends much more on the details of the numerical simulations. In particular we show that adding force noise to the numerical data set can significantly alter the topological features in the data. We conclude that both fNR and topological metrics are useful measures to consider when quantifying the state of a granular system.

7.5. Miscellaneous

7.5.1. On Order Types of Random Point Sets

Participant: Marc Glisse.

In collaboration with Olivier Devillers and Xavier Goaoc (Inria team Gamble) and Philippe Duchon (LaBRI, Université de Bordeaux).

Let P be a set of n random points chosen uniformly in the unit square. In this paper [41], we examine the typical resolution of the order type of P. First, we show that with high probability, P can be rounded to the grid of step $\frac{1}{n^{3+\epsilon}}$ without changing its order type. Second, we study algorithms for determining the order type of a point set in terms of the number of coordinate bits they require to know. We give an algorithm that requires on average $4n \log_2 n + O(n)$ bits to determine the order type of P, and show that any algorithm requires at least $4n \log_2 n - O(n \log \log n)$ bits. Both results extend to more general models of random point sets.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- Collaboration with Sysnav, a French SME with world leading expertise in navigation and geopositioning in extreme environments, on TDA, geometric approaches and machine learning for the analysis of movements of pedestrians and patients equipped with inetial sensors (CIFRE PhD of Bertrand Beaufils).
- Research collaboration with Fujitsu on the development of new TDA methods and tools for Machine learning and Artificial Intelligence (started in Dec 2017).

8.2. Bilateral Grants with Industry

- DATASHAPE and Sysnav have been selected for the ANR/DGA Challenge MALIN (funding: 700 kEuros) on pedestrian motion reconstruction in severe environments (without GPS access).

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

9.1.1.1. ANR ASPAG

Participant: Marc Glisse.

- Acronym : ASPAG.

- Type : ANR blanc.

- Title : Analysis and Probabilistic Simulations of Geometric Algorithms.

- Coordinator : Olivier Devillers (équipe Inria Gamble).

- Duration : 4 years from January 2018 to December 2021.

- Others Partners: Inria Gamble, LPSM, LABRI, Université de Rouen, IECL, Université du Littoral Côte d'Opale, Telecom ParisTech, Université Paris X (Modal'X), LAMA, Université de Poitiers, Université de Bourgogne.

- Abstract:

The analysis and processing of geometric data has become routine in a variety of human activities ranging from computer-aided design in manufacturing to the tracking of animal trajectories in ecology or geographic information systems in GPS navigation devices. Geometric algorithms and probabilistic geometric models are crucial to the treatment of all this geometric data, yet the current available knowledge is in various ways much too limited: many models are far from matching real data, and the analyses are not always relevant in practical contexts. One of the reasons for this state of affairs is that the breadth of expertise required is spread among different scientific communities (computational geometry, analysis of algorithms and stochastic geometry) that historically had very little interaction. The Aspag project brings together experts of these communities to address the problem of geometric data. We will more specifically work on the following three interdependent directions.

(1) Dependent point sets: One of the main issues of most models is the core assumption that the data points are independent and follow the same underlying distribution. Although this may be relevant in some contexts, the independence assumption is too strong for many applications.

(2) Simulation of geometric structures: The phenomena studied in (1) involve intricate random geometric structures subject to new models or constraints. A natural first step would be to build up our understanding and identify plausible conjectures through simulation. Perhaps surprisingly, the tools for an effective simulation of such complex geometric systems still need to be developed.

(3) Understanding geometric algorithms: the analysis of algorithm is an essential step in assessing the strengths and weaknesses of algorithmic principles, and is crucial to guide the choices made when designing a complex data processing pipeline. Any analysis must strike a balance between realism and tractability; the current analyses of many geometric algorithms are notoriously unrealistic. Aside from the purely scientific objectives, one of the main goals of Aspag is to bring the communities closer in the long term. As a consequence, the funding of the project is crucial to ensure that the members of the consortium will be able to interact on a very regular basis, a necessary condition for significant progress on the above challenges.

- See also: https://members.loria.fr/Olivier.Devillers/aspag/

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

9.2.1.1. GUDHI

Title: Algorithmic Foundations of Geometry Understanding in Higher Dimensions

Programm: FP7

Type: ERC

Duration: February 2014 - January 2019

Coordinator: Inria

Inria contact: Jean-Daniel Boissonnat.

The central goal of this proposal is to settle the algorithmic foundations of geometry understanding in dimensions higher than 3. We coin the term geometry understanding to encompass a collection of tasks including the computer representation and the approximation of geometric structures, and the inference of geometric or topological properties of sampled shapes. The need to understand geometric structures is ubiquitous in science and has become an essential part of scientific computing and data analysis. Geometry understanding is by no means limited to three dimensions. Many applications in physics, biology, and engineering require a keen understanding of the geometry of a variety of higher dimensional spaces to capture concise information from the underlying often highly nonlinear structure of data. Our approach is complementary to manifold learning techniques and aims at developing an effective theory for geometric and topological data analysis. To reach these objectives, the guiding principle will be to foster a symbiotic relationship between theory and practice, and to address fundamental research issues along three parallel advancing fronts. We will simultaneously develop mathematical approaches providing theoretical guarantees, effective algorithms that are amenable to theoretical analysis and rigorous experimental validation, and perennial software development. We will undertake the development of a high-quality open source software platform to implement the most important geometric data structures and algorithms at the heart of geometry understanding in higher dimensions. The platform will be a unique vehicle towards researchers from other fields and will serve as a basis for groundbreaking advances in scientific computing and data analysis.

9.3. International Research Visitors

9.3.1. Visits of International Scientists

- Wolfgang Polonik, UC Davis, California. Sept. and Oct. 2018. Statistical aspects of persistent homology.
- Arijit Ghosh, Indian Statistical Institute, Kolkata, India (December 2018)
- Ramsay Dyer, Berkeley Publishing (December 2018)

9.3.1.1. Internships

- Shreya Arya, BITS Pilani University, India, August-July 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of Organizing Committees

- F. Chazal co-organised the Tutorial “Machine Learning on Evolutionary Computation” at the IEEE World Congress on Computational Intelligence (WCCI), Rio de Janeiro, July 2018.
- J-D. Boissonnat was a member of the organization committee of the International Conference on Curves and Surfaces, Arcachon, July 2018.
- S. Oudot organized the mini-symposium on topological data analysis and learning at the International Conference on Curves and Surfaces, Arcachon, July 2018.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- S. Oudot was a PC member of the International Symposium on Computational Geometry (SoCG), Budapest, Hungary, June 2018.
- David Cohen-Steiner was a PC member of the Symposium on Geometry Processing (SGP), Paris, France, July 2018, and of Shape Modeling International (SMI), Lisbon, Portugal, June 2018.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

Jean-Daniel Boissonnat is a member of the Editorial Board of *Journal of the ACM*, *Discrete and Computational Geometry*, *International Journal on Computational Geometry and Applications*.

Frédéric Chazal is a member of the Editorial Board of *SIAM Journal on Imaging Sciences*, *Discrete and Computational Geometry (Springer)*, *Graphical Models (Elsevier), and Journal of Applied and Computational Topology (Springer)*.

Steve Oudot is a member of the Editorial Board of *Journal of Computational Geometry*.

10.1.4. Invited Talks

Frédéric Chazal, Abel Symposium, Geiranger, Norway, June 2018.

Frédéric Chazal, Colloquium de Mathématiques, Math Dept. Amiens, October 2018.

Frédéric Chazal, AI Research Center at National Cheng-Kung University, Taiwan, May 2018.

Frédéric Chazal, National Center for High-performance Computing, Taiwan, May 2018.

Jean-Daniel Boissonnat, Hamilton Mathematics Institute, Trinity College, Dublin, Ireland, June 2018.

Steve Oudot, Workshop “Topological Data Analysis meets Symplectic Topology”, Tel Aviv, Israel, May 2018.

Steve Oudot, Abel Symposium, Geiranger, Norway, June 2018.

Steve Oudot, Banff workshop on multiparameter persistence, Oaxaca, Mexico, August 2018.

Steve Oudot, ICERM, Brown University, Providence, USA, August 2018.

Steve Oudot, workshop on structural inference in high-dimensional models, Moscow, Russia, September 2018.

Clément Maria, Einstein workshop on Geometric and Topological Combinatorics, Freie Universität, Berlin, Germany, October 2018.

10.1.5. Leadership within the Scientific Community

Frédéric Chazal is co-responsible, with S. Arlot (Paris-Sud Univ.), of the “programme Maths-STIC” of the Labex Fondation Mathématique Jacques Hadamard (FMJH).

Frédéric Chazal has been a member of the Scientific council of the french “Agence pour les Mathematiques en Interaction avec l’Entreprise et la Societe (AMIES)” until Dec. 2018.

Frédéric Chazal is a member of the “Comité de pilotage” of the SIGMA group at SMAI.

Steve Oudot is co-organizing the monthly seminar on combinatorial and computational geometry at Institut Henri Poincaré.

Steve Oudot is co-head (with Luca Castelli-Aleardi) of the GT Géométrie Algorithmique within the GdR Informatique Mathématique.

Steve Oudot is a member of the program committee of the DataIA convergence institute.

10.1.6. Scientific Expertise

- Consulting collaboration for IFPEN to explore potential applications of TDA (from February 2018 to Dec. 2018).

10.1.7. Research Administration

Frédéric Chazal is a member of the Équipe de Direction at Inria Saclay.

Marc Glisse, responsable Raweb pour DataShape

Steve Oudot is vice-president of the Commission Scientifique at Inria Saclay.

Clément Maria is a member of the CDT at Inria Sophia Antipolis-Méditerranée.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Master: Frédéric Chazal and Quentin Mérigot, Analyse Topologique des Données, 30h eq-TD, Université Paris-Sud, France.

Master: Jean-Daniel Boissonnat and Marc Glisse, Computational Geometry Learning, 36h eq-TD, M2, MPRI, France.

Master: Frédéric Cazals and Frédéric Chazal, Geometric Methods for Data Analysis, 30h eq-TD, M1, École Centrale Paris, France.

Master: Frédéric Chazal and Julien Tierny, Topological Data Analysis, 38h eq-TD, M2, Mathématiques, Vision, Apprentissage (MVA), ENS Paris-Saclay, France.

Master: Steve Oudot, Topological data analysis, 45h eq-TD, M1, École polytechnique, France.

Master: Steve Oudot, Data Analysis: geometry and topology in arbitrary dimensions, 24h eq-TD, M2, graduate program in Artificial Intelligence & Advanced Visual Computing, École polytechnique, France.

Undergrad-Master: Steve Oudot, preparatory course for international programming contests, 54h eq-TD, L3/M1, École polytechnique, France.

Summer School on topological data analysis and persistent homology: Steve Oudot, advanced topics, 6h eq-TD, Trento, Italy, June 2018.

Summer School on geometric data: Frédéric Chazal and Marc Glisse, Introduction to Topological Data Analysis, 9h eq-TD, Fréjus, Sept. 2018.

Winter School on Computational Geometry, Amirkabir University of Technology, Tehran, Iran. Course on Delaunay Triangulation of Manifolds, March 2018.

10.2.2. Supervision

PhD : Claire Brécheteau, Statistical aspects of distance-like functions , Defended on September 2018, Frédéric Chazal (co-advised by Pascal Massart).

PhD in progress: Bertrand Beaufils, Méthodes topologiques et apprentissage statistique pour l'actimétrie du piéton à partir de données de mouvement, started November 2016, Frédéric Chazal (co-advised by Bertrand Michel).

PhD: Jérémy Cochoy, Decomposition and stability of multidimensional persistence modules, Defended on December 10, 2018, Steve Oudot.

PhD in progress: Yitchzak Solomon, Inverse problems in topological data analysis, started September 1st, 2016, Steve Oudot (co-advised by Jeff Brock, Brown University).

PhD in progress: Nicolas Berkouk, Categorification of topological graph structures, started November 1st, 2016, Steve Oudot.

PhD in progress: Théo Lacombe, Statistics for persistence diagrams using optimal transport, started October 1st, 2017, Steve Oudot.

PhD in progress: Alba Chiara de Vitis, Concentration of measure and clustering, Jean-Daniel Boissonnat and David Cohen-Steiner.

PhD in progress: Siargey Kachanovich, Manifold reconstruction in higher dimensions, Jean-Daniel Boissonnat.

PhD in progress: Siddharth Pritam, Approximation algorithms in Computational Topology, Jean-Daniel Boissonnat.

PhD in progress: Raphaël Tinarrage, Persistence and stability of nerves in measured metric spaces for Topological Data Analysis, started September 1st, 2017, Frédéric Chazal and Marc Glisse.

PhD in progress: Vincent Divol, statistical aspects of TDA, started September 1st, 2017, Frédéric Chazal (co-advised by Pascal Massart).

PhD in progress: Owen Rouillé, September 2018, co-advised by C. Maria and J-D. Boissonnat.

10.2.3. Juries

J-D. Boissonnat was a member of the committee for the HDR defense of Aurélien Alvarez (Université d'Orléans).

F. Chazal was a member of the PhD defense committee of Jisu Kim (Carnegie Mellon University, advisors: A. Rinaldo and L. Wasserman), Claire Brécheteau (Université Paris-Saclay, advisors: F. Chazal and P. Massart), Hariprasad Kannan (Centrale-Supelec, advisor: N. Paragios), Dorian Nognen (Ecole Polytechnique, advisor: M. Ovsjanikov).

S. Oudot was a member of the Ph.D. defence committee of Tim Ophelders (T.U. Eindhoven, advisors: Bettina Speckmann and Kevin Buchin).

10.3. Popularization

10.3.1. Interventions

- Frédéric Chazal: Fujitsu Forum, "Topological Data Analysis: from academic success to industrial innovation", Tokyo, Japan, May 2018.
- Frédéric Chazal: "TDA and AI for biomedical applications", Kaohsiung MEdical Technology Expo, Taiwan, May 2018.

11. Bibliography

Major publications by the team in recent years

[1] D. ATTALI, U. BAUER, O. DEVILLERS, M. GLISSE, A. LIEUTIER. *Homological Reconstruction and Simplification in R3*, in "Computational Geometry", 2014 [*DOI :* 10.1016/J.COMGEO.2014.08.010], https://hal.archives-ouvertes.fr/hal-01132440

[2] J.-D. BOISSONNAT, R. DYER, A. GHOSH. *Delaunay Triangulation of Manifolds*, in "Foundations of Computational Mathematics", 2017, vol. 45, 38 p. [*DOI :* 10.1007/S10208-017-9344-1], https://hal.inria.fr/hal-01509888

[3] J.-D. BOISSONNAT, R. DYER, A. GHOSH, S. Y. OUDOT. *Only distances are required to reconstruct submanifolds*, in "Computational Geometry", 2017, vol. 66, pp. 32 - 67 [*DOI :* 10.1016/J.COMGEO.2017.08.001], https://hal.inria.fr/hal-01583086

[4] J.-D. BOISSONNAT, K. C. SRIKANTA, S. TAVENAS. *Building Efficient and Compact Data Structures for Simplicial Complexe*, in "Algorithmica", September 2016 [*DOI :* 10.1007/S00453-016-0207-Y], https://hal.inria.fr/hal-01364648

[5] F. CHAZAL, D. COHEN-STEINER, A. LIEUTIER. *A Sampling Theory for Compact Sets in Euclidean Space*, in "Discrete Comput. Geom.", 2009, vol. 41, n^{o} 3, pp. 461–479, http://dx.doi.org/10.1007/s00454-009-9144-8

[6] F. CHAZAL, D. COHEN-STEINER, Q. MÉRIGOT. *Geometric Inference for Measures based on Distance Functions*, in "Foundations of Computational Mathematics", 2011, vol. 11, n^{o} 6, pp. 733-751, RR-6930 [*DOI :* 10.1007/S10208-011-9098-0], http://hal.inria.fr/inria-00383685

[7] F. CHAZAL, S. Y. OUDOT, M. GLISSE, V. DE SILVA. *The Structure and Stability of Persistence Modules*, SpringerBriefs in Mathematics, Springer Verlag, 2016, VII, 116 p. , https://hal.inria.fr/hal-01330678

[8] L. J. GUIBAS, S. Y. OUDOT, P. SKRABA, F. CHAZAL. *Persistence-Based Clustering in Riemannian Manifolds*, in "Journal of the ACM", November 2013, vol. 60, n^{o} 6, 38 p. , http://hal.archives-ouvertes.fr/hal-00923563

[9] M. MANDAD, D. COHEN-STEINER, L. KOBBELT, P. ALLIEZ, M. DESBRUN. *Variance-Minimizing Transport Plans for Inter-surface Mapping*, in "ACM Transactions on Graphics", 2017, vol. 36, 14 p. [*DOI :* 10.1145/3072959.3073671], https://hal.inria.fr/hal-01519006

[10] S. Y. OUDOT. *Persistence Theory: From Quiver Representations to Data Analysis*, Mathematical Surveys and Monographs, American Mathematical Society, 2015, n^{o} 209, 218 p. , https://hal.inria.fr/hal-01247501

Publications of the year

Articles in International Peer-Reviewed Journals

[11] S. ARYA, G. DA FONSECA, D. MOUNT. *Approximate Polytope Membership Queries*, in "SIAM Journal on Computing", January 2018, vol. 47, n^{o} 1, pp. 1 - 51 [*DOI :* 10.1137/16M1061096], https://hal.archives-ouvertes.fr/hal-01890054

[12] J.-D. BOISSONNAT, C. KARTHIK. *An Efficient Representation for Filtrations of Simplicial Complexes*, in "ACM Transactions on Algorithms", September 2018, vol. 14, https://hal.inria.fr/hal-01883836

[13] T. BONIS, S. OUDOT. *A Fuzzy Clustering Algorithm for the Mode-Seeking Framework*, in "Pattern Recognition Letters", 2018, https://arxiv.org/abs/1406.7130 - Submitted to Pattern Recognition Letters [*DOI :* 10.1016/J.PATREC.2017.11.019], https://hal.inria.fr/hal-01111854

[14] D. BREMNER, O. DEVILLERS, M. GLISSE, S. LAZARD, G. LIOTTA, T. MCHEDLIDZE, G. MOROZ, S. WHITESIDES, S. WISMATH. *Monotone Simultaneous Paths Embeddings in* $\mathbb{R}^d$, in "Discrete Mathematics

and Theoretical Computer Science", January 2018, vol. 20, n^o 1, pp. 1-11 [*DOI :* 10.23638/DMTCS-20-1-1], https://hal.inria.fr/hal-01529154

[15] M. CARRIERE, B. MICHEL, S. Y. OUDOT. *Statistical analysis and parameter selection for Mapper*, in "Journal of Machine Learning Research", 2018, https://hal.archives-ouvertes.fr/hal-01633106

[16] F. CHAZAL, B. FASY, F. LECCI, B. MICHEL, A. RINALDO, L. WASSERMAN. *Robust Topological Inference: Distance To a Measure and Kernel Distance*, in "Journal of Machine Learning Research", 2018, vol. 18, n^o 159, 40 p. , https://arxiv.org/abs/1412.7197v1 - To appear, forthcoming, https://hal.inria.fr/hal-01232217

[17] C. COLONNELLO, M. KRAMÁR. *Dynamics of silo deformation under granular discharge*, in "pre physical review E", 2018, https://hal.inria.fr/hal-01849279

[18] J. DIJKSMAN, L. KOVALCINOVA, J. REN, R. BEHRINGER, M. KRAMÁR, K. MISCHAIKOW, L. KONDIC. *Characterizing Granular Networks Using Topological Metrics*, in "Physical Review E ", 2018, vol. 97, 042903 p. , https://hal.inria.fr/hal-01757079

[19] P. DLOTKO, T. WANNER. *Rigorous cubical approximation and persistent homology of continuous functions*, in "Computers and Mathematics with Applications", March 2018, https://hal.archives-ouvertes.fr/hal-01706695

[20] L. KANARI, P. DLOTKO, M. SCOLAMIERO, R. LEVI, J. SHILLCOCK, K. HESS, H. MARKRAM. *Quantifying topological invariants of neuronal morphologies*, in "Neuroinformatics", January 2018, https://arxiv.org/abs/1603.08432 - 10 pages, 5 figures, conference or other essential info, https://hal.archives-ouvertes.fr/hal-01706793

[21] C. SORZANO, A. JIMÉNEZ, J. MOTA, J. VILAS, D. MALUENDA, M. MARTÍNEZ, E. RAMÍREZ-APORTELA, T. MAJTNER, J. SEGURA, R. SÁNCHEZ-GARCÍA, Y. RANCEL, L. DEL CAÑO, P. CONESA, R. MELERO, S. JONIC, J. VARGAS, F. CAZALS, Z. FREYBERG, J. KRIEGER, I. BAHAR, R. MARABINI, J. CARAZO. *Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy*, in "Acta crystallographica. Section F, Structural biology communications", January 2019, vol. 75, n^o 1, pp. 19-32, https://hal.inria.fr/hal-01968170

International Conferences with Proceedings

[22] S. ARYA, G. D. DA FONSECA, D. M. MOUNT. *Approximate Convex Intersection Detection with Applications to Width and Minkowski Sums*, in "ESA 2018 - European Symposium on Algorithms", Helsinki, Finland, 26th Annual European Symposium on Algorithms proceedings, August 2018, https://arxiv.org/abs/1807.00484 [*DOI :* 10.4230/LIPICS.ESA.2018.3], https://hal.archives-ouvertes.fr/hal-01890039

[23] B. BEAUFILS, F. CHAZAL, M. GRELET, B. MICHEL. *Activity recognition from stride detection: a machine learning approach based on geometric patterns and trajectory reconstruction*, in "IPIN 2018 - 9th International Conference on Indoor Positioning and Indoor Navigation", Nantes, France, Proceedings IPIN 2018 - 9th International Conference on Indoor Positioning and Indoor Navigation, September 2018, https://hal.archives-ouvertes.fr/hal-01864467

[24] J.-D. BOISSONNAT, K. DUTTA, A. GHOSH, S. KOLAY. *Tight Kernels for Covering and Hitting: Point Hyperplane Cover and Polynomial Point Hitting Set*, in "LATIN 2018 - 13th Latin American Theoretical INformatics Symposium", Buenos Aires, Argentina, April 2018, https://hal.inria.fr/hal-01669884

[25] J.-D. BOISSONNAT, R. DYER, A. GHOSH, M. WINTRAECKEN. *Local Criteria for Triangulation of Manifolds*, in "International Symposium on Computational Geometry", Budapest, Hungary, June 2018 [*DOI :* 10.4230/LIPICS.SOCG.2018.9], https://hal.inria.fr/hal-01801616

[26] J.-D. BOISSONNAT, A. LIEUTIER, M. WINTRAECKEN. *The reach, metric distortion, geodesic convexity and the variation of tangent spaces* , in "SoCG 2018 - 34th International Symposium on Computational Geometry", Budapest, Hungary, June 2018, pp. 1-14 [*DOI :* 10.4230/LIPICS.SOCG.2018], https://hal.inria.fr/hal-01801667

[27] J.-D. BOISSONNAT, S. PRITAM, D. PAREEK. *Strong Collapse for Persistence*, in "ESA 2018 - 26th Annual European Symposium on Algorithms", Helsinki, Finland, August 2018, pp. 67:1–67:13 [*DOI :* 10.4230/LIPICS], https://hal.inria.fr/hal-01886165

[28] F. CHAZAL, V. DIVOL. *The density of expected persistence diagrams and its kernel based estimation*, in "Symposium of Computational Geometry (SoCG 2018)", Budapest, Hungary, June 2018, Extended version of a paper to appear in the proceedings of the Symposium of Computational Geometry 2018, https://hal.archives-ouvertes.fr/hal-01716181

[29] D. COHEN-STEINER, W. KONG, C. SOHLER, G. VALIANT. *Approximating the Spectrum of a Graph*, in "KDD 2018 - Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining", London, United Kingdom, August 2018, https://arxiv.org/abs/1712.01725 , https://hal.inria.fr/hal-01661199

[30] T. LACOMBE, M. CUTURI, S. OUDOT. *Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport*, in "NIPS", Montreal, Canada, 2018, https://hal.inria.fr/hal-01966674

Scientific Books (or Scientific Book chapters)

[31] J.-D. BOISSONNAT, F. CHAZAL, M. YVINEC. *Geometric and Topological Inference*, Cambridge University Press, 2018, Cambridge Texts in Applied Mathematics, https://hal.inria.fr/hal-01615863

Other Publications

[32] E. AAMARI, J. KIM, F. CHAZAL, B. MICHEL, A. RINALDO, L. WASSERMAN. *Estimating the Reach of a Manifold*, January 2018, https://arxiv.org/abs/1705.04565 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01521955

[33] H. ANAI, F. CHAZAL, M. GLISSE, Y. IKE, H. INAKOSHI, R. TINARRAGE, Y. UMEDA. *DTM-based filtrations*, November 2018, https://arxiv.org/abs/1811.04757 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01919562

[34] S. ARYA, J.-D. BOISSONNAT, K. DUTTA. *Persistent Homology with Dimensionality Reduction: k-Distance vs Gaussian Kernels*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01950051

[35] J.-D. BOISSONNAT, O. DEVILLERS, K. DUTTA, M. GLISSE. *Randomized incremental construction of Delaunay triangulations of nice point sets*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01950119

[36] J.-D. BOISSONNAT, S. KACHANOVICH, M. WINTRAECKEN. *Triangulating submanifolds: An elementary and quantified version of Whitney's method*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01950149

[37] J.-D. BOISSONNAT, S. PRITAM. *Computing Persistent Homology of Flag Complexes via Strong Collapses*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01950074

[38] C. BRÉCHETEAU, A. FISCHER, C. LEVRARD. *Robust Bregman Clustering*, December 2018, https://arxiv.org/abs/1812.04356 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01948051

[39] C. BRÉCHETEAU, C. LEVRARD. *The k-PDTM : a coreset for robust geometric inference*, January 2018, https://arxiv.org/abs/1801.10346 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01694542

[40] D. COHEN-STEINER, A. C. DE VITIS. *Spectral Properties of Radial Kernels and Clustering in High Dimensions*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01969956

[41] O. DEVILLERS, P. DUCHON, M. GLISSE, X. GOAOC. *On Order Types of Random Point Sets*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01962093

[42] V. DIVOL, W. POLONIK. *On the choice of weight functions for linear representations of persistence diagrams*, July 2018, https://arxiv.org/abs/1807.03678 - working paper or preprint, https://hal.inria.fr/hal-01833660

[43] S. HARKER, M. KRAMÁR, R. LEVANGER, K. MISCHAIKOW. *A Comparison Framework for Interleaved Persistence Modules*, April 2018, working paper or preprint, https://hal.inria.fr/hal-01757092

[44] M. KERBER, M. LESNICK, S. OUDOT. *Exact computation of the matching distance on 2-parameter persistence modules*, 2018, working paper or preprint, https://hal.inria.fr/hal-01966666

[45] C. MARIA, H. SCHREIBER. *Discrete Morse Theory for Computing Zigzag Persistence*, January 2019, https://arxiv.org/abs/1807.05172 - working paper or preprint, https://hal.inria.fr/hal-01971682

[46] S. OUDOT, E. SOLOMON. *Barcode Embeddings for Metric Graphs*, February 2018, https://arxiv.org/abs/1712.03630 - working paper or preprint, https://hal.inria.fr/hal-01708780

[47] S. OUDOT, E. SOLOMON. *Inverse Problems in Topological Persistence: a Survey*, 2018, https://arxiv.org/abs/1810.10813 - working paper or preprint, https://hal.inria.fr/hal-01966676