

Paris-Saclay, le 9 novembre 2017

Monsieur OUDOT Steve

Inria Saclay

1 rue Honoré d'Estienne d'Orves, Bâtiment Alan

Turing, Campus de l'Ecole Polytechnique

91120 Palaiseau

Objet : **Soutenance de doctorat**

Monsieur,

**La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique**

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

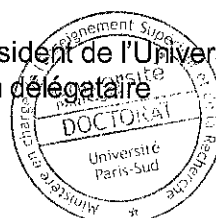
46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur CARLSSON Gunnar
Stanford University
Office 383-L
450 Serra Mall
Bldg. 380
CA 94305-2125 Stanford

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

**Télécom ParisTech Bâtiment C,
46 rue Barrault,
75013 Paris**

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur VERT Jean-Philippe
Centre for Computational Biology
Office V315
Mines ParisTech
60 boulevard Saint-Michel
75006 Paris

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

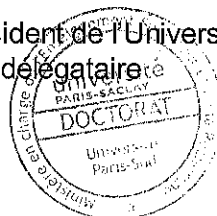
46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur MAIRAL Julien
Inria Grenoble
655, avenue de l'Europe
38330 Montbonnot Saint Martin

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur GOAOC Xavier
Université Paris Est
5, boulevard Descartes - Champs sur Marne
77454 Marne-la-Vallée

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

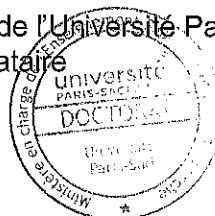
46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur BAUER Ulrich
TUM Munich
Boltzmannstraße 3
D-85747 Garching

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



Paris-Saclay, le 9 novembre 2017

Monsieur SCHOENAUER Marc

Inria Saclay

1 rue Honoré d'Estienne d'Orves, Bâtiment Alan

Turing, Campus de l'Ecole Polytechnique

91120 Palaiseau

Objet : **Soutenance de doctorat**

Monsieur,

La soutenance de doctorat de Monsieur Mathieu CARRIERE,
en Informatique

préparée à l'université Paris-Sud

au sein de l'école doctorale Sciences et Technologies de l'Information et de la
Communication

sur le sujet :

**« Sur les propriétés métriques et statistiques des descripteurs topologiques pour les
données géométriques »**

aura lieu :

LE MARDI 21 NOVEMBRE 2017 à 8h00

à

C47

Télécom ParisTech Bâtiment C,

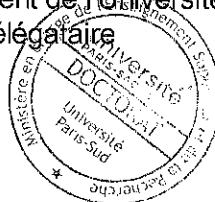
46 rue Barrault,

75013 Paris

Vous êtes invité à y participer, en tant que membre du jury. Vous trouverez ci-joints les
rapports des rapporteurs.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Le président de l'Université Paris Saclay,
ou son délégué



STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-2125

GUNNAR CARLSSON
ANN AND BILL SWINDELLS PROFESSOR
DEPARTMENT OF MATHEMATICS
(650) 723-2224
gunnar@math.stanford.edu

October 24, 2017

**Report on *Sur les propri  s m  triques et statistiques des descripteurs*
topologiques pour les donn  es g  om  triques by M. Carri  re**

This dissertation studies a very interesting and important problem within topological data analysis. I will first summarize the context and then discuss the contributions made in the present dissertation.

Topological data analysis is a relatively new area of inquiry within mathematics and computer science. It has been underdevelopment over the last 15-20 years. Its goal is to provide methods for understanding the “shape” of data sets, suitably interpreted. A useful way to define a shape on a finite set of points is through a *metric* or *distance function*, which is typically encoding information about similarity of data points. There have been two separate directions of inquiry within the subject.

- **Measuring shape:** The field of algebraic topology within pure mathematics defines algebraic invariants of shapes (perhaps very high dimensional) that capture the presence of gross features, such as loops, spheres, etc.. There are variants that capture other things, such as the presence of flares etc.. One of the achievements in the subject has been to extend these methods (which apply to situations where we have complete information about the shapes that we are dealing with) to situations where we only have information about a sample from an underlying space. This extension is referred to as *persistent homology*. While the output of standard

homology consists of numbers (so-called Betti numbers), which roughly count the instances of certain patterns occurring in the shape), the new methodology produces outputs that are called bar codes, or persistence diagrams. They consist of collections of intervals, each capturing a particular feature, which an extent that is indicative of the scale of the feature. This more continuous representation allows one to infer the presence of these features in an underlying space from which the points have been sampled. Many variants of these constructions have been developed, and have been found to be extremely useful for many different scientific problems.

- **Representing shape:** It is also useful to deal directly with the shape of the data set. A natural class of representations of shapes consists of *simpli-
cial complexes*, which are collections of points, edges, triangles, tetrahedra, and higher dimensional analogues, with suitable intersection properties. The simplest examples are graphs, in the computer science sense. One of the main threads in ordinary topology is the approximation of spaces by such complexes. Indeed, this is essential for the definition and computation of the homological invariants mentioned above. A family of methods which has long been used in data analysis is described as *cluster analysis*. In this case, the complexes have only points, with no edges or higher dimensional objects. There are many constructors for complexes based on distance functions, including Čech, Vietoris-Rips, alpha shapes, witness complexes, etc.. However, many of these create very high dimensional complexes, which is very problematic for dealing directly with the shape. There is one method, Mapper, which explicitly controls the dimension of the complex, and it has been found to be extremely useful for many applied problems. It has been used to discover the decomposition of type 2 diabetes and asthma into smaller different disease forms, which will lead to much more targeted methods for treatment. It has also been used to map the state space of subjects undergoing infection by a particular disease,

allowing for different approaches to the understanding of the function of the immune system. There are many other examples. Once constructed, the Mapper model complex can be laid out on the screen using standard layout algorithms, and it is possible to interact with it in numerous ways, including the selection of subsets, the encoding by coloring of function values on the data set, finding the variables that characterize the various regions in the data set, etc.

One of the problems of the Mapper construction is that it involves parameters and can sometimes exhibit instability with respect to the choice of these parameters. This dissertation is a step toward the solution of this problem. It uses methods from both threads above. Specifically, it does four things.

1. It constructs a pseudometric on objects closely related to the Mapper construction, namely *Reeb graphs*. Mapper (in its one-dimensional form) is a discretized version of Reeb graphs. The pseudometric is obtained by computing persistence diagrams for the Reeb graph, and using known distance functions on the set of persistence diagrams. The reason it is a pseudometric is that it is quite possible for somewhat different Reeb graphs to have identical persistence diagrams. However, it is understood that large changes in the Reeb graph will be captured in the persistence diagrams.
2. The relationship between Mapper (applied to a space) and its idealized version, a Reeb graph, is analyzed, and this analysis enables the extension of the pseudometric for Reeb graphs to Mapper outputs. This allows for the proof of a certain kind of stability theorem for Mapper, which is a very important contribution. This is an excellent result which has appeared in the proceedings of the Symposium on Computational Geometry, the premier computer science conference for computational geometry.
3. Prove a convergence theorem for Mapper applied to point clouds within

a space to the Mapper construction applied to the full space. This “completes” the circle, and permits the statistical analysis of the construction, in terms of various kinds of features.

4. Analyze kernel methods for the space of persistence diagrams. This is important in that it will make possible better inference methods for the study of persistence diagrams, and therefore of shapes.

This dissertation is an outstanding piece of work. It is clearly of theoretical importance, and I think is also extremely likely that it will be of importance in the applied world. It may over time suggest methods for modifying the construction so that it will be stable, and will give quantifiable methods for understanding the presence of features complex data sets. These are both very desirable goals, in that they will allow users to more quickly gain confidence in the results they observe. I strongly recommend its acceptance.

Sincerely yours,

A handwritten signature in black ink, appearing to read "Gunnar Carlsson". The signature is fluid and cursive, with a long horizontal stroke at the end.

Gunnar Carlsson
Ann and Bill Swindells Professor, Emeritus
Department of Mathematics
Stanford University

Rapport de Julien Mairal et Jean-Philippe Vert sur la thèse de Mathieu Carrière

Sur les propriétés métriques et statistiques des descripteurs topologiques pour les données géométriques

La thèse de Mathieu Carrière étudie une approche originale pluri-disciplinaire consistant à représenter les données par leurs attributs topologiques. A la frontière de la géométrie computationnelle, des statistiques, et de l'analyse de données, il s'agit d'un sujet conceptuellement élégant, mathématiquement très intéressant, mais qui est traditionnellement peu abordé par la communauté d'apprentissage statistique, dont les interactions avec la communauté de géométrie computationnelle sont très rares. Dans sa thèse, Mathieu Carrière réussit le tour de force de contribuer de façon significative dans les deux disciplines et d'établir des ponts importants, qui pourront potentiellement inspirer une génération nouvelle de travaux.

En géométrie computationnelle, ses contributions mathématiques sur la stabilité des mappers, ainsi que son travail sur les métriques permettant de comparer des graphes de Reeb ont ainsi donné lieu à trois publications dont l'une au Journal of Foundations of Computational Mathematics et deux à la conférence SoCG. A l'intersection des deux disciplines, une méthode de représentation des diagrammes de persistance pour la classification de formes 3D a été publiée dans la conférence SGP en 2015. Récemment, les travaux de Mathieu Carrière ont aussi été reconnus par la communauté d'apprentissage statistique, avec une publication à la conférence hautement sélective ICML (il s'agit d'une des deux conférences majeures d'apprentissage automatique, qui joue un rôle aussi important que les journaux internationaux). Cette dernière contribution est par ailleurs extrêmement originale. Bien que moins technique que les précédentes, elle fait un lien entre l'analyse de données topologique, le transport optimal, et les méthodes à noyaux en apprentissage statistique, ce qui témoigne d'un spectre de compétences très large et d'une grande maturité scientifique.

Les chapitres introductifs 1 et 2 présentent respectivement une description des enjeux, des outils, et des contributions de façon non-formelle et intuitive, et une introduction des outils mathématiques utilisés dans la thèse (diagrammes de persistance, graphes de Reeb, Mapper). Le chapitre 1 fait preuve d'une grande pédagogie et constitue un point d'entrée clair et intuitif pour des membres de la communauté d'apprentissage statistique. Le chapitre 2 introduit les mêmes outils de façon abstraite et mathématique. Le chapitre est alors beaucoup plus aride, mais introduit en environ 25 pages, ce qui est usuellement introduit en plusieurs chapitres d'ouvrages.

**CENTRE DE RECHERCHE
GRENOBLE - RHÔNE-ALPES**

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex France
Tél. : +33 (0)4 76 61 52 00
Fax : +33 (0)4 76 61 52 52

ANTENNE Inria LYON LA DOUA

Bâtiment CEI-1
66 Boulevard Niels Bohr
69603 Villeurbanne France
Tél. : +33 (0)4 72 43 74 90
Fax : +33 (0)4 72 43 74 99

Le chapitre 3 apporte des justifications théoriques à l'utilisation de la distance "bottleneck" d_b pour comparer des graphes de Reeb. Celle-ci représente ces graphes par des diagrammes de persistance étendus et n'est donc qu'une pseudo-métrique (deux graphes de Reeb différents pouvant avoir le même diagramme de persistance étendu), ce qui rend discutable son utilisation pour des tâches discriminatives. Cependant, la métrique d_b a l'avantage d'être calculable en pratique, contrairement à d'autres métriques pour les graphes de Reeb, telles que la distortion fonctionnelle d_{FD} ou la distance de Gromov-Hausdorff d_{GH} . Le chapitre présente alors deux résultats positifs concernant la distance d_b . Tout d'abord, *localement*, d_b est équivalente aux métriques mentionnées précédemment. Par ailleurs, la métrique induite intrinsèque est *globalement* équivalente aux métriques intrinsèques induites par d_{FD} et d_{GH} . Bien que n'étant pas très bien placés pour juger de l'importance de cette contribution dans la littérature de géométrie computationnelle, nous avons trouvé les résultats très intéressants, ainsi que les arguments avancés compréhensibles et cohérents. Les deux résultats précédents nous semblent être des étapes significatives permettant de mieux comprendre la structure locale de l'espace des graphes de Reeb. Le chapitre se conclut par une liste de questions ouvertes intéressantes comme l'existence ou non de plus courts chemins dans l'espace des graphes de Reeb, qui permettrait d'utiliser en pratique la métrique intrinsèque induite par d_b .

Le chapitre 4 se focalise ensuite sur l'outil Mapper, qui est dérivé des graphes de Reeb, et qui a connu un certain nombre de succès applicatifs. Le chapitre étend la distance de bottleneck aux diagrammes de persistance d'une variante du Mapper (multinerve Mapper). L'intérêt est ensuite de montrer que cette pseudo-métrique est stable (non-expansive par rapport à la norme ℓ_∞). Étant donné que Mapper peut être vu comme une version pratique "pixelisée" du graphe de Reeb (le chapitre introduit par ailleurs un résultat de convergence du Mapper vers le graphe de Reeb), cette contribution permet de mieux comprendre le succès du Mapper, qui a démontré des qualités importantes dans des applications.

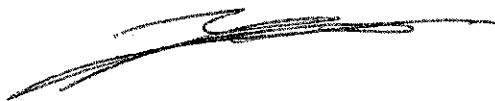
Le chapitre 5 s'intéresse à des aspects statistiques moins abstraits que les deux chapitres précédents, permettant de caractériser le taux de convergence de Mapper vers le graphe de Reeb, lorsque le Mapper est calculé sur un nuage de points échantillonné à partir d'une variété de dimension d . Ce chapitre permet de faire le lien entre les outils de géométrie computationnelle décrits dans les chapitres précédents, et des applications réelles où les données sont échantillonnées. Le résultat principal montre que le mapper converge en espérance vers le graphe de Reeb, en distance de bottleneck, avec un taux borné par $O((\log(n)/n)^{1/d})$ pour des fonctions régulières. Une borne inférieure est aussi calculée, montrant que le taux est optimal d'un point de vue minimax. Ce résultat a deux implications majeures. Tout d'abord, il montre que la dimension de la variété joue un rôle critique (ce qui n'est intuitivement pas surprenant, mais une justification théorique est tout à fait bienvenue). En second lieu, il donne

aussi lieu à une heuristique pratique permettant de choisir une couverture spécifique pour le Mapper. Une application intéressante de l'analyse statistique permet aussi de calculer des diagrammes de persistance incluant des intervalles de confiance.

Enfin, le chapitre 6 présente des distances pratiques permettant de comparer des diagrammes de persistance, et de les utiliser dans des applications d'apprentissage automatique. La contribution de ce chapitre est particulièrement originale, et permet d'établir un point entre géométrie computationnelle, transport optimal, et méthodes à noyaux. La distance obtenue utilise le noyau "sliced Wasserstein" et il est démontré, que sous certaines hypothèses légères, la distance induite par le noyau est équivalente à la distance de Wasserstein pour les diagrammes de persistance (bien que la distance de Wasserstein ne soit pas une métrique Hilbertienne). Le chapitre aborde alors des questions pratiques computationnelles, permettant d'approcher cette distance, et enfin des applications convaincantes de la distance pour des tâches de classification. Il s'agit ainsi d'un chapitre très complet, abordant des questions de modélisation (design d'une métrique adaptée aux données, d'un point de vue topologique), des questions théoriques (stabilité de la métrique), et aussi pratiques avec plusieurs applications, qui ne manqueront pas d'éveiller l'intérêt de la communauté d'apprentissage statistique.

Pour conclure, Matthieu Carrière a effectué une thèse excellente, et a contribué de façon significative à plusieurs domaines scientifiques, ce qui est remarquable. Il a su résoudre certains problèmes ouverts théoriques demandant une analyse mathématique fine, proposer de nouveaux modèles et outils, et a su appliquer (de façon parcimonieuse) ses recherches à des problèmes pratiques. En conséquence, nous émettons un avis extrêmement favorable à l'autorisation de soutenance de ses travaux, en vue de l'obtention de son diplôme de doctorat de l'université Paris Saclay.

Julien Mairal



Jean-Philippe Vert



**CENTRE DE RECHERCHE
GRENOBLE - RHÔNE-ALPES**

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex France
Tél. : +33 (0)4 76 61 52 00
Fax : +33 (0)4 76 61 52 52

www.inria.fr

ANTENNE Inria LYON LA DOUA

Bâtiment CCI-1
66 Boulevard Niels Bohr
69603 Villeurbanne France
Tél. : +33 (0)4 72 43 74 90
Fax : +33 (0)4 72 43 74 99