Analyse topologique de données out-of-core

Julien Tierny

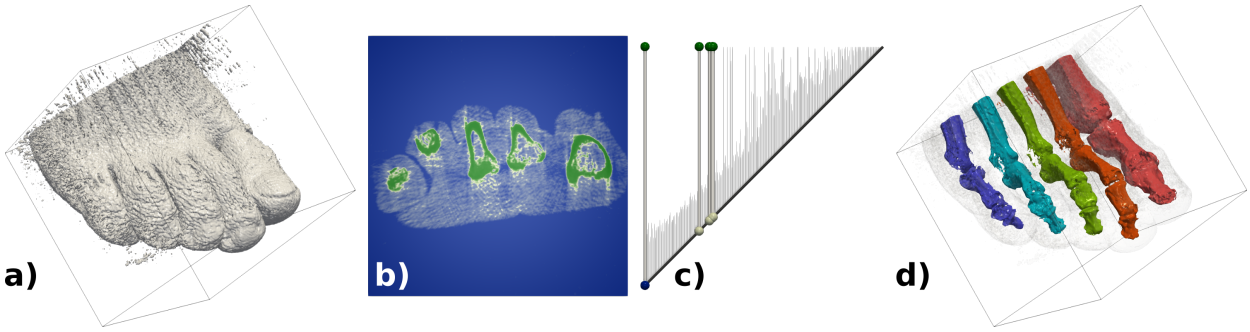

Fig. 1. Le sujet en une image – L'homologie persistante est un outil théorique puissant, qui permet en pratique d'introduire une mesure de bruit sur les structures topologiques au sein de données complexes. Cette mesure de bruit, appelée *persistance*, permet de visualiser et de mesurer des structures topologiques à plusieurs échelles d'importance et d'extraire efficacement et avec précision les structures d'intérêt dans un jeu de données. La persistance est souvent associée au *diagramme de persistance* (c), qui donne une représentation visuelle de la distribution des structures topologiques en fonction de leur plage de valeur dans les données. Ces diagrammes, grâce à leur stabilité [9], jouent un rôle central en analyse topologique de données: ils constituent en effet une représentation réduite des données particulièrement pertinente, qui capture les structures les plus importantes des données. Dans cet exemple (scan CT d'un pied humain – a: isosurface, b: coupe dans le volume), l'utilisateur a sélectionné les cinq structures les plus persistantes sur le diagramme (c: barres blanches pleines), aboutissant ainsi à une segmentation automatique des os des différents doigts du pied (d) [1,5]. Comme illustré par cet exemple, les diagrammes de persistance jouent donc un rôle important pour l'analyse interactive de données complexes. Cependant, leur calcul pratique sur des jeux de données de taille actuellement réaliste (typiquement 2048^3) pose encore des difficultés, notamment en terme d'empreinte mémoire. Dans ce stage, nous souhaitons répondre à ce problème en concevant une approche *out-of-core*, permettant de calculer le diagramme de persistance de jeux de données trop volumineux pour rentrer entièrement en mémoire.

1 Contexte

L'analyse topologique de données (TDA) [7, 10, 30, 37, 38] est une discipline à cheval entre informatique et mathématiques appliquées qui propose d'analyser des données complexes au vu de leur structure, de leur topologie [28]. Elle connait un essor important depuis quelques années, dû principalement à ses succès récents en *data science* [30,38] et en *machine learning* [6], ainsi qu'aux développements open-source récents la mettant en oeuvre [12,41].

Parmi les différents outils d'analyse développés en TDA (comme le graphe de Reeb [17,31,43], le complexe de Morse-Smale [14,20,26], etc...), l'homologie persistante [10,11] est un outil fondamental qui propose de mesurer l'importance des structures topologiques (composantes connexes, cycles, cavités, etc.) selon leur *durée de vie* (leur plage de valeur dans les données). Cette théorie permet d'introduire une mesure de bruit sur les structures topologiques, appelée *persistance*, dont la stabilité a été démontrée d'un point de vue théorique [9], et qui permet en pratique de distinguer avec efficacité et précision les structures d'intérêt du bruit (voir Fig. 1). L'efficacité pratique de l'homologie persistante a été documentée dans de nombreuses applications, comme en imagerie médicale [1,5], en biologie cellulaire [21], en mécanique des fluides [8,16,23,34,39], en physique des matériaux [13,22,25], en combustion [3,4,19,24], en chimie moléculaire [2,15], en astrophysique [33,36], en traitement de surfaces [40,42,44,45], en compression [35] ou encore en monitoring de simulations numériques haute-performance [32].

• *Julien Tierny is with Sorbonne Université, CNRS, LIP6 UMR 7606, France. E-mails: julien.tierny@sorbonne-universite.fr.*

2 Problème scientifique

Les données considérées sont typiquement représentées sous la forme d'une fonction scalaire linéaire par morceaux $f : \mathcal{M} \rightarrow \mathbb{R}$, associant une valeur réelle à chaque sommet d'une triangulation $\mathcal{M}$, qui représente un objet géométrique 2D ou 3D. En pratique, f représente dans les applications des niveaux de concentrations [13], des potentiels [15], des intensités [1], des températures [4], etc. Les sous-ensembles de niveau $f^{-1}_{-\infty}(i)$ sont définis comme la pré-image de l'intervalle ouvert $]-\infty, i[$ sur $\mathcal{M}$. Simplement, il s'agit de l'ensemble des points de l'objet au dessous d'une certaine valeur i. Quand i augmente, $f^{-1}_{-\infty}(i)$ change de topologie en un nombre fini de configurations: ses nombres de Betti [46] (nombres de composantes connexes, de cycles indépendants, de cavités, etc...) changent sur des points singuliers, appelés points critiques. Chaque structure topologique de $f^{-1}_{-\infty}(i)$ est donc créée sur un premier point critique à une valeur i, puis détruite sur un second point critique à une valeur $j > i$. Le diagramme de persistance $\mathscr{D}(f)$ [9, 11] (Fig. 1c) est une représentation graphique de ce processus, où chaque classe d'homologie persistante (chaque structure topologique) est représentée par une barre verticale pour laquelle la cordonnée en abscisse correspond à la valeur i et les extrémités en ordonnées correspondent à i et j. La *persistance* de la classe est donnée par $|j-i|$. Dans ce diagramme, le bruit topologique apparait donc sous la forme de petites barres, proche de la diagonale ($|j-i| \rightarrow 0$), voir [37].

Comme illustré Fig. 1, le diagramme de persistance est particulièrement utile pour visualiser la distribution des structures topologiques en fonction de leur importance. Il permet également de segmenter assez directement les données pour extraire les structures les plus importantes (ici les cinq os du pied) à des fins d'analyse quantitative [1,5].

Cependant, le calcul du diagramme de persistance [18] pose encore des difficultés en terme d'empreinte mémoire pour des jeux de données

de taille réaliste. Pour le jeu de données illustré Fig. 1, d'une résolution de 256^3 sommets, le calcul requiert en pratique 5 Go de mémoire vive. Or, de nos jours, une taille réaliste pour des jeux de données de ce type est plus de 2048^3 (512 fois plus gros, nécessitant potentiellement plus de 2 To de mémoire vive). Bien que des optimisations soient envisageables en terme d'implémentation, les ordres de grandeur en jeux sont tels qu'une refonte profonde de l'algorithme est nécessaire.

L'objectif de ce stage est donc de définir un nouvel algorithme pour le calcul *out-of-core* de diagrammes de persistance, spécialisé pour traiter des jeux de données trop volumineux pour rentrer entièrement en mémoire. La stratégie *divide-and-conquer* est une piste naturelle à la démarche générale de calcul *out-of-core*. Elle consiste à charger les données petit bout (*chunk*) par petit bout, à effectuer des calculs locaux sur ces données partielles, puis à intégrer l'ensemble des calculs locaux pour obtenir le résultat final global. Parmi les pistes possibles, nous nous inspirerons des approches *divide-and-conquer* pour le calcul de *merge tree* [27, 29].

3 Perspectives

Ce stage est proposé dans l'optique d'une poursuite en thèse de doctorat sur le thème de l'analyse topologique de données. Cette thèse peut être financée dans le contexte d'un partenariat CIFRE avec Kitware, société majeure dans la data-science et le logiciel open-source (CMake [48], CDash [47], VTK [52], ITK [49], ParaView [51]). Elle peut aussi être financée dans le cadre d'un autre partenariat CIFRE, avec Total, entreprise également intéressée par l'analyse topologique pour le traitement de ses données de simulation et d'acquisition.

Ce stage peut également déboucher sur un poste d'ingénieur de recherche (CDD) ouvert à Sorbonne Université pour le développement de la bibliothèque TTK [41] (financement européen H2020-FET).

De manière plus générale, ce stage et sa possible poursuite en thèse apporteront un bagage de compétences scientifiques et techniques pointues et recherchées dans le domaine de la *data science* et de l'analyse et de la visualisation interactive de données scientifiques (TDA, TTK [41], ParaView [51]). Il constitue donc une expérience fortement valorisable pour accéder à des fonctions R&D sur ces thèmes, dans le monde académique comme industriel (Kitware, EDF, Total, CEA, etc.).

4 Organisation du stage

Le stage pourra se dérouler selon les étapes suivantes:

1. Etudier la bibliographie existante sur:
   - l'analyse topologique de données [10, 37];
   - le calcul de diagramme de persistance [18];
   - les approches *divide-and-conquer* pour le calcul de *merge tree* [27, 29].
2. Imaginer et mettre en oeuvre un algorithme *out-of-core* pour le calcul de diagrammes de persistances;
3. Valider l'approche d'un point de vue expérimental sur une variété de jeux de données pratiques provenant de divers contextes applicatifs.

Les programmes d'expérimentation seront écrits en C++, sous la forme de modules pour la plate-forme open-source d'analyse topologique de données *"Topology ToolKit"* (TTK) [41] (intégrée à ParaView [51]).

Le stage peut durer de 16 à 24 semaines, selon les disponibilités du stagiaire. Il s'agit d'un stage rémunéré (rémunération académique standard, environ 500 euros par mois).

5 Profil

Nous recherchons un(e) étudiant(e) très motivé(e)! Curiosité, ouverture d'esprit, créativité, et ténacité sont les aptitudes de caractère que nous recherchons. Ce stage s'adresse aux étudiants en dernière année de master en informatique ou mathématiques appliquées (et domaines connexes) ou aux étudiants en dernière année d'école d'ingénieurs. Le stagiaire devra être à l'aise avec la programmation en C++, ou motivé pour le devenir. Un intérêt pour la 3D, la géométrie, la topologie et plus généralement pour les mathématiques et l'informatique est requis.

6 Lieu

Ce stage aura lieu au laboratoire d'informatique (LIP6) de Sorbonne Université, en plein coeur de Paris (arrêt Jussieu, lignes 7 et 10). Il sera encadré par Julien Tierny, chercheur au CNRS, expert en analyse topologique de données pour la visualisation et l'analyse de données scientifiques (http://lip6.fr/Julien.Tierny).

7 Candidatures

Nous invitons les candidat(e)s à nous faire parvenir leur lettre de candidature accompagnée d'un CV mis à jour à Julien Tierny (julien.tierny@sorbonne-universite.fr). Nous vous encourageons à nous contacter par email pour toute question ou pour discuter davantage du sujet.

References

[1] A. Bock, H. Doraiswamy, A. Summers, and C. Silva. Topoangler: Interactive topology-based extraction of fishes. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017.

[2] R. A. Boto, J. C. Garcia, J. Tierny, and J.-P. Piquemal. Interpretation of the reduced density gradient. *Molecular Physics*, 2016.

[3] P. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. A topological framework for the interactive exploration of large scale turbulent combustion. In *Proc. of IEEE eScience*, 2009.

[4] P. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. Interactive exploration and analysis of large scale simulations using topology-based data segmentation. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[5] H. Carr, J. Snoeyink, and M. van de Panne. Simplifying flexible isosurfaces using local geometric measures. In *IEEE VIS*, 2004.

[6] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 2013.

[7] F. Chazal and J. Tierny. Topological data analysis, online class. `http://lip6.fr/Julien.Tierny/topologicalDataAnalysisClass.html`.

[8] F. Chen, H. Obermaier, H. Hagen, B. Hamann, J. Tierny, and V. Pascucci. Topology analysis of time-dependent multi-fluid data using the reeb graph. *Computer Aided Geometric Design*, 2013.

[9] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Symp. on Comp. Geom.*, 2005.

[10] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.

[11] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Disc. Compu. Geom.*, 2002.

[12] G. Favelier, C. Gueunet, A. Gyulassy, J. Jomier, J. Levine, J. Lukasczyk, D. Sakurai, M. Soler, J. Tierny, W. Usher, and Q. Wu. Topological data analysis made easy with the Topology ToolKit. In *Proc. of IEEE VIS Tutorials*, 2018. `https://topology-tool-kit.github.io/ieeeVisTutorial.html`.

[13] G. Favelier, C. Gueunet, and J. Tierny. Visualizing ensembles of viscous fingers. In *IEEE SciVis Contest*, 2016.

[14] R. Forman. A user's guide to discrete Morse theory. *Adv. in Math.*, 1998.

[15] D. Guenther, R. Alvarez-Boto, J. Contreras-Garcia, J.-P. Piquemal, and J. Tierny. Characterizing molecular interactions in chemical systems. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2014.

[16] D. Guenther, J. Salmon, and J. Tierny. Mandatory critical points of 2D uncertain scalar fields. *Computer Graphics Forum (Proc. of EuroVis)*, 2014.

[17] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Contour forests: Fast multi-threaded augmented contour trees. In *IEEE LDAV*, 2016.

[18] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Task-based Augmented Merge Trees with Fibonacci Heaps,. In *IEEE LDAV*, 2017.

[19] A. Gyulassy, P. Bremer, R. Grout, H. Kolla, J. Chen, and V. Pascucci. Stability of dissipation elements: A case study in combustion. *Computer Graphics Forum (Proc. of EuroVis)*, 2014.

[20] A. Gyulassy, P. T. Bremer, B. Hamann, and V. Pascucci. A practical approach to morse-smale complex computation: Scalability and generality. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2008.

[21] A. Gyulassy, D. Guenther, J. A. Levine, J. Tierny, and V. Pascucci. Conforming morse-smale complexes. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2014.

[22] A. Gyulassy, A. Knoll, K. Lau, B. Wang, P. Bremer, M. Papka, L. A. Curtiss, and V. Pascucci. Interstitial and interlayer ion diffusion geometry extraction in graphitic nanosphere battery materials. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2015.

[23] J. Kasten, J. Reininghaus, I. Hotz, and H. Hege. Two-dimensional time-dependent vortex regions based on the acceleration magnitude. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[24] D. E. Laney, P. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2006.

[25] J. Lukasczyk, G. Aldrich, M. Steptoe, G. Favelier, C. Gueunet, J. Tierny, R. Maciejewski, B. Hamann, and H. Leitte. Viscous fingering: A topological visual analytic approach. In *Physical Modeling for Virtual Manufacturing Systems and Processes*, 2017.

[26] J. Milnor. *Morse Theory*. Princeton U. Press, 1963.

[27] D. Morozov and G. Weber. Distributed merge trees. In *ACM Symposium on Principles and Practice of Parallel Programming*, 2013.

[28] MyWhyU. A humorous look at the topology of curved space. `https://www.youtube.com/watch?v=p2ofJPh2yMw&list=PL09E9E697F585A58C`.

[29] V. Pascucci and K. Cole-McLaughlin. Parallel computation of the topology of level sets. *Algorithmica*, 2003.

[30] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny. *Topological Methods in Data Analysis and Visualization: Theory, Algorithms and Applications*. Springer, 2010.

[31] G. Reeb. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *Acad. des Sci.*, 1946.

[32] E. Santos, J. Tierny, A. Khan, B. Grimm, L. Lins, J. Freire, V. Pascucci, C. Silva, S. Klasky, R. Barreto, and N. Podhorszki. Enabling advanced visualization tools in a web-based simulation monitoring system. In *Proc. of IEEE eScience*, 2009.

[33] N. Shivashankar, P. Pranav, V. Natarajan, R. van de Weygaert, E. P. Bos, and S. Rieder. Felix: A topology based framework for visual exploration of cosmic filaments. *IEEE Transactions on Visualization and Computer Graphics*, 2016. `http://vgl.serc.iisc.ernet.in/felix/index.html`.

[34] M. Soler, M. Plainchault, B. Conche, and J. Tierny. Lifted Wasserstein matcher for fast and robust topology tracking. In *Proc. of IEEE Symposium on Large Data Analysis and Visualization*, 2018.

[35] M. Soler, M. Plainchault, B. Conche, and J. Tierny. Topologically controlled lossy compression. In *Proc. of IEEE PacificViz*, 2018.

[36] T. Sousbie. The persistent cosmic web and its filamentary structure: Theory and implementations. *Royal Astronomical Society*, 2011. `http://www2.iap.fr/users/sousbie/web/html/indexd41d.html`.

[37] J. Tierny. Introduction to topological data analysis. `https://hal.archives-ouvertes.fr/cel-01581941/file/manuscript.pdf`.

[38] J. Tierny. *Topological Data Analysis for Scientific Visualization*. Springer, 2018.

[39] J. Tierny and H. Carr. Jacobi fiber surfaces for bivariate Reeb space computation. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2016.

[40] J. Tierny, J. Daniels, L. G. Nonato, V. Pascucci, and C. Silva. Interactive quadrangulation with Reeb atlases and connectivity textures. *IEEE Transactions on Visualization and Computer Graphics*, 2012.

[41] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017. `https://topology-tool-kit.github.io/`.

[42] J. Tierny, D. Guenther, and V. Pascucci. Optimal general simplification of scalar fields on surfaces. In *Topological and Statistical Methods for Complex Data*. Springer, 2014.

[43] J. Tierny, A. Gyulassy, E. Simon, and V. Pascucci. Loop surgery for volumetric meshes: Reeb graphs reduced to contour trees. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2009.

[44] J. Tierny and V. Pascucci. Generalized topological simplification of scalar fields on surfaces. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2012.

[45] A. Vintescu, F. Dupont, G. Lavoué, P. Memari, and J. Tierny. Conformal factor persistence for fast hierarchical cone extraction. In *Eurographics (short papers)*, 2017.

[46] Wikipedia. Betti numbers. `https://en.wikipedia.org/wiki/Betti_number`.

[47] Wikipedia. CDash. `https://www.cdash.org/`.

[48] Wikipedia. CMake. `https://en.wikipedia.org/wiki/CMake`.

[49] Wikipedia. ITK. `https://en.wikipedia.org/wiki/Insight_Segmentation_and_Registration_Toolkit`.

[50] Wikipedia. Kitware Inc. `https://en.wikipedia.org/wiki/Kitware`.

[51] Wikipedia. ParaView. `https://en.wikipedia.org/wiki/ParaView`.

[52] Wikipedia. VTK. `https://en.wikipedia.org/wiki/VTK`.