Patania et al. *EPJ Data Science* (2017) 6:7
DOI 10.1140/epjds/s13688-017-0104-x

EPJ Data Science
a SpringerOpen Journal

COMMENTARY Open Access

Topological analysis of data

Alice Patania[1,2*], Francesco Vaccarino[1,2] and Giovanni Petri[1]

*Correspondence:
alice.patania@isi.it
[1]ISI Foundation, via Alassio 11c, Turin, Italy
[2]Dipartimento di Scienze Matematiche "G.L. Lagrange", Politecnico di Torino, corso Duca degli Abruzzi 24, Turin, Italy

Abstract

Propelled by a fast evolving landscape of techniques and datasets, data science is growing rapidly. Against this background, topological data analysis (TDA) has carved itself a niche for the analysis of datasets that present complex interactions and rich structures. Its distinctive feature, topology, allows TDA to detect, quantify and compare the mesoscopic structures of data, while also providing a language able to encode interactions beyond networks. Here we briefly present the TDA paradigm and some applications, in order to highlight its relevance to the data science community.

Keywords: topological data analysis; simplicial complexes; persistent homology

1 Introduction

The science of data is constantly changing. Classic [1] and emerging [2] machine learning tools and network-theoretic descriptions [3, 4] are opening windows on the intimate details of systems as diverse as individual behavioural and mobility patterns [5], the human consciousness and language [6], the control and robustness of biological systems [7, 8], and even quantum dynamics [9].

While network approaches can describe the fabric of relations between agents in a system, or molecules in an organism, they are constrained in their descriptive power to pairwise interactions (i.e. edges), which might not always be justified when focusing on phenomena that involve group dynamics (e.g. scientific collaboration, genetic pathways) or higher-order descriptions (e.g. viral evolution, molecule folding). Similarly, machine learning tools are extremely efficient in classifying and segmenting datasets into consistent patterns or clusters. However, at times they find themselves challenged when asked to produce an organic description of the interactions among a system's components, and they often suffer from the curse of dimensionality due to their underlying geometric formalization. Over the last decade, a set of new techniques for data analysis, based on a set-theoretic formalism, has been gaining traction. The set-theoretic foundation makes them topological in nature (and hence geometry-independent), and so they have come to be collectively referred to as Topological Data Analysis (TDA). The reason for the growing interest in TDA is its capacity to capture the large- and mesoscale shape of datasets via their algebraic-topological structure [10]: on the one hand, TDA allows one to enrich network descriptions with higher order ones (i.e. many-body interactions); on the other, it adds also a notion of organization, or *shape*, to the descriptions obtained from traditional classification techniques. In this note, we aim to introduce TDA, some of its successful ap-

plications to real data analysis and why we believe it to be important for the data science community as a set of goggles complementary to the existing ones.

2 Why topological data analysis?

The novelty of TDA is that it studies the shape of topological spaces at the mesoscopic scale by going beyond standard measures defined on data points' pairs. This is done by moving from networks to simplicial complexes. The latter are obtained from elementary objects, called simplices, built from such simple polyhedra as points, line segments, triangles, tetrahedra, and their higher dimensional analogues, glued together along their faces.

Simplicial complexes were first introduced in 1895 by Poincaré in his seminal work "Analysis Situs" [11] as a simplicial decomposition (triangulation) of a manifold. They have been since used to store in discrete form key information on a topological space and to transform complicated topological problems into more familiar algebraic ones with the introduction of simplicial homology (we refer to Aleksandrov [12] for a beautiful account of the birth of combinatorial topology). Being the fundamental method in combinatorial topology [13], the use of simplicial complexes is not new in science, as for example, they are the secret behind every 3D rendering and image recognition software [14]. They have however taken a new life with the emergence of TDA techniques [15, 16].

In this arena, simplicial complexes constitute the choice representation of many-body interactions of complex systems. In fact, by glueing together simplices of different sizes and composition, one is able to describe varied, heterogeneous and changing interactions. The resulting simplicial complexes efficiently summarise the shape of the underlying datasets and yield mesoscopic information about how simplices coordinate with one another across intermediate and large scales within the complexes. TDA summaries can be read out from the simplicial complexes directly, or by studying the patterns of holes in all dimensions that define their shapes (via the corresponding homology groups) [17]. These summaries are both informative and guaranteed to be robust to perturbations [18]: in particular, they do not vary under changes in coordinates or under deformations of the individual samples, which makes them parsimonious descriptions of arbitrary datasets. The possibility to represent complex interactions of any order, together with the robustness to missing and corrupted data quality issues common many real datasets, is the rationale behind the growing number of applications in biology [19], neuroscience [20, 21], social sciences [22, 23], physics [24, 25], quantum computation [26], and nanotechnologies [27].

3 TDA in practice

In everyday applications, the TDA tool-kit consists essentially of two main techniques: topological simplification (via the Mapper algorithm [28, 29]), and persistent homology [30, 31].

3.1 Topological simplification

The aim of topological approaches to data is to produce sparser readable summaries of complex datasets. Mapper, introduced by Singh et al. [29], is the main tool for this type of direct data exploration. It produces a topological skeleton of a dataset (akin to a balls-and-sticks representation) by slicing the data-space in overlapping slabs according to case-specific quantities and performing local clustering within each. The resulting clusters are then linked together to recover a simplified yet (provably) complete picture of the overall

topology, which however provides more structured information than standard techniques, such as PCA, MDS, or clustering techniques alone could.

The fact that the clustering is performed locally in each slice makes this approach computationally convenient as it can be easily be computed in parallel. It is therefore a good tool for the analysis of large-scale datasets which can be used in a framework of big data analysis such as the Google's MapReduce paradigm [32]. Indeed, Mapper is most well-known for the discovery of a new subtype of breast cancer [33] from genetic data, but it has also found successful applications in other biomedical studies [34], e.g. the identification of diabetes subtypes [35] and of different pulmonary conditions [36], as well as in industrial [37] and commercial applications (e.g. Ayasdi).

3.2 Persistent homology

Despite Mapper's facility of interpretation and computational advantages, it does not explicitly yield quantitative insights that allow for direct comparison within and across datasets. Persistent homology is able to do that. Unlike Mapper, it does not compress the data. Rather, it encodes data in a simplicial filtration, a series of progressively finer simplicial complexes. This filtration is then analysed to build a multi-scale low-dimensional summary that tracks the lifespan and evolution of connected components, holes and high dimensional voids along the sequence of simplicial complexes. That is, it identifies and quantifies the different kinds of 'empty space' embedded in the data, which implicitly make up the dataset's shape. This grants the possibility to obtain insights into unique mesoscale structures otherwise invisible to standard analytical tools, which in turn motivates an ever growing application of persistent homology across fields, such as biology [38, 39], social science [22, 31] and neuroscience [20]. For example, in biology, Chan et al. [40] showed that viral genetic recombination is captured better by homological invariant than standard phylogenetic trees across a number of diseases. In the social sciences, Bajardi et al. [41] used homological features to characterise the correlations between socio-economic indicators and spatial structure of migrant communities in Milan. Persistent homology has found its widest application so far in the study of structural and functional brain connectivity where the range of scales and the complexity of the systems seem to benefit the most from a persistent homology description (e.g. [42–44]).

Unfortunately, these features are somewhat less intuitive than those Mapper provides. However, statistical mechanical methods can give an important contribution to their interpretation. For example, Lord et al. [45] and Verovsek et al. [46] projected the results of persistent homology to simpler representation, i.e. lower dimensional scaffolds or skeletons providing localized information and making topological features amenable to network techniques. Also, Kahle [47] and Courtney et al. [48] have taken the first steps in constructing minimal topological random null models in order to provide a notion of which topological features should be considered significant and which ones noise.

4 Conclusion

TDA is still very much developing as a branch of data science. It provides a new paradigm, based on algebraic topology, for how we think about data, and has obtained its first successes. However, there still are many challenges to be met to fully exploit its potential. The most pressing one is the computational scalability of persistent homology, which currently prevents large-scale applications. Although the classic algorithms for topological

features extraction have a memory and time complexity which is polynomial on the number of simplices, and the latter is exponential on the number of vertices. While the last five years have witnessed significant advances in parallel algorithms [49–51] and simplicial reduction schemes [52–55], new implementations and algorithmic improvements are paramount. A second issue is that of the localization of homological features within a given dataset. This is crucial if we want to pinpoint the effect of specific topological features (e.g. a certain disconnectivity pattern in brain regions, or a structural hole among scientific or professional collaborators), and to leverage it beyond classification purposes. Steps in this direction have been taken [46, 56], but more work is required to make this information easily accessible and interpretable. Tightly related to the previous point is the need to strengthen the link between machine learning and TDA. A few initial works have proposed ways to make persistent homology's outputs directly amenable to standard machine learning techniques (typically via kernel approaches [57–59]). However, there is still a wide gap to be bridged to obtain a productive integration of TDA's novel perspective within the full machine-learning framework. Finally, at the community level, it is also necessary to help practitioners coming from outside the TDA community to discover and adopt these techniques. Efforts in this direction (see for example the introduction by [60], the development of shared TDA R libraries [61, 62]) are undergoing, but there is still ample space for contributions from the machine learning, mathematical, and complex networks communities.

Acknowledgements
GP and AP acknowledge the support of the ADnD project by Compagnia San Paolo. FV acknowledges the support of the Laboratorio Lagrange project by Fondazione CRT.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AP, GP and FV designed and wrote the manuscript. All authors read and approved the final manuscript.

Publisher's Note

Received: 15 May 2017 Accepted: 29 May 2017 Published online: 06 June 2017

References

1. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
2. Bengio Y, Goodfellow IJ, Courville A (2015) Deep learning. Nature 521:436-444
3. Newman M, Barabasi A-L, Watts DJ (2006) The structure and dynamics of networks. Princeton studies in complexity. Princeton University Press, Princeton
4. Zanin M, Papo D, Sousa PA, Menasalvas E, Nicchi A, Kubik E, Boccaletti S (2016) Combining complex networks and data mining: why and how. Phys Rep 635:1-44
5. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782
6. Bassett DS, Sporns O (2017) Network neuroscience. Nat Neurosci 20(3):353-364
7. Gao J, Barzel B, Barabási A-L (2016) Universal resilience patterns in complex networks. Nature 530(7590):307-312
8. Gates AJ, Rocha LM (2016) Control of complex networks requires both structure and dynamics. Sci Rep 6:24456
9. Carleo G, Troyer M (2017) Solving the quantum many-body problem with artificial neural networks. Science 355(6325):602-606. doi:10.1126/science.aag2302
10. Carlsson G, Zomorodian A, Collins A, Guibas LJ (2005) Persistence barcodes for shapes. Int J Shape Model 11(02):149-187
11. Poincaré H (1895) Analysis situs. J Éc Polytech 1:1-123
12. Aleksandrov PS (1972) Poincaré and topology. Russ Math Surv 27(1):157-168
13. Steenrod SE-N, Eilenberg S (1952) Foundations of algebraic topology. Princeton University Press, Princeton
14. Kovalevsky VA (1989) Finite topology as applied to image analysis. Comput Vis Graph Image Process 46(2):141-161
15. Edelsbrunner H, Letscher D, Zomorodian A (2002) Topological persistence and simplification. Discrete Comput Geom 28(4):511-533

16. Carlsson G (2009) Topology and data. Bull Am Math Soc 46(2):255-308
17. Ghrist R (2008) Barcodes: the persistent topology of data. Bull Am Math Soc 45(1):61-75. doi:10.1090/S0273-0979-07-01191-3
18. Cohen-Steiner D, Edelsbrunner H, Harer J (2007) Stability of persistence diagrams. Discrete Comput Geom 37(1):103-120
19. Chan JM, Carlsson G, Rabadan R (2013) Topology of viral evolution. Proc Natl Acad Sci 110(46):18566-18571
20. Giusti C, Pastalkova E, Curto C, Itskov V (2015) Clique topology reveals intrinsic geometric structure in neural correlations. Proc Natl Acad Sci 112(44):13455-13460. doi:10.1073/pnas.1506407112
21. Yoo J, Kim EY, Ahn YM, Ye JC (2016) Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. J Neurosci Methods 267:1-13
22. Carstens C, Horadam K (2013) Persistent homology of collaboration networks. Math Probl Eng 2013:815035
23. Schebesch KB, Stecking RW (2017) Topological data analysis for extracting hidden features of client data. In: Operations research proceedings 2015. Springer, Berlin, pp 483-489
24. Hiraoka Y, Nakamura T, Hirata A, Escolar EG, Matsue K, Nishiura Y (2016) Hierarchical structures of amorphous solids characterized by persistent homology. Proc Natl Acad Sci 113(26):7035-7040. doi:10.1073/pnas.1520877113
25. Donato I, Gori M, Pettini M, Petri G, De Nigris S, Franzosi R, Vaccarino F (2016) Persistent homology analysis of phase transitions. Phys Rev E 93(5):052138
26. Lloyd S, Garnerone S, Zanardi P (2016) Quantum algorithms for topological and geometric analysis of data. Nat Commun 7:10138
27. Nakamura T, Hiraoka Y, Hirata A, Escolar EG, Nishiura Y (2015) Persistent homology and many-body atomic structure for medium-range order in the glass. Nanotechnology 26(30):304001
28. Singh G, Memoli F, Carlsson G (1991) Mapper: a topological mapping tool for point cloud data. In: Eurographics symposium on point-based graphics
29. Singh G, Mémoli F, Carlsson GE (2007) Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: SPBG, pp 91-100
30. Ghrist R (2008) Barcodes: the persistent topology of data. Bull Am Math Soc 45(1):61-75
31. Petri G, Scolamiero M, Donato I, Vaccarino F (2013) Topological strata of weighted complex networks. PLoS ONE 8(6):66506
32. Dean J, Ghemawat S (2008) Mapreduce: simplified data processing on large clusters. Commun ACM 51(1):107-113
33. Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc Natl Acad Sci 108(17):7265-7270
34. Lum P, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013) Extracting insights from the shape of complex data using topology. Sci Rep 3:1236
35. Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT (2015) Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 7(311):311ra174
36. Rucco M, Merelli E, Herman D, Ramanan D, Petrossian T, Falsetti L, Nitti C, Salvi A (2015) Using topological data analysis for diagnosis pulmonary embolism. J Theor Appl Comput Sci 9(1):41-55
37. Guo W, Banerjee AG (2016) Toward automated prediction of manufacturing productivity based on feature selection using topological data analysis. In: Assembly and manufacturing (ISAM), 2016 IEEE international symposium on. IEEE Press, New York, pp 31-36
38. Xia K, Wei G-W (2014) Persistent homology analysis of protein structure, flexibility, and folding. Int J Numer Methods Biomed Eng 30(8):814-844
39. Mamuye A, Rucco M (2016) Persistent homology on rna secondary structure space. In: Proceedings of the 9th EAI international conference on bio-inspired information and communications technologies (formerly BIONETICS). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Gent, pp 189-192
40. Chan JM, Carlsson G, Rabadan R (2013) Topology of viral evolution. Proc Natl Acad Sci 110(46):18566-18571
41. Bajardi P, Delfino M, Panisson A, Petri G, Tizzoni M (2015) Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Sci 4(1):3
42. Giusti C, Pastalkova E, Curto C, Itskov V (2015) Clique topology reveals intrinsic geometric structure in neural correlations. Proc Natl Acad Sci 112(44):13455-13460
43. Sizemore A, Giusti C, Betzel RF, Bassett DS (2016) Closures and cavities in the human connectome. Preprint. arXiv:1608.03520
44. Bendich P, Marron J, Miller E, Pieloch A, Skwerer S (2016) Persistent homology analysis of brain artery trees. Ann Appl Stat 10(1):198
45. Lord L-D, Expert P, Fernandes HM, Petri G, Van Hartevelt TJ, Vaccarino F, Deco G, Turkheimer F, Kringelbach ML (2016) Insights into brain architectures from the homological scaffolds of functional connectivity networks. Front Syst Neurosci 10:85
46. Verovsek SK, Kurlin V, Lesnik D (2017) The higher-dimensional skeletonization problem. Preprint. arXiv:1701.08395
47. Kahle M (2014) Sharp vanishing thresholds for cohomology of random flag complexes. Ann Math (2) 179(3):1085-1107. doi:10.4007/annals.2014.179.3.5
48. Courtney OT, Bianconi G (2016) Generalized network structures: the configuration model and the canonical ensemble of simplicial complexes. Preprint. arXiv:1602.04110
49. Edelsbrunner H, Harer J (2010) Computational topology - an introduction. Am. Math. Soc., Providence, p 1241
50. Bauer U, Kerber M, Reininghaus J, Wagner H (2014) Phat–persistent homology algorithms toolbox. In: International congress on mathematical software. Springer, Berlin, pp 137-143
51. Bauer U, Kerber M, Reininghaus J (2014) Clear and compress: computing persistent homology in chunks. In: Mathematics and visualization. Springer, Berlin, pp 103-117
52. Busaryev O, Cabello S, Chen C, Dey TK, Wang Y (2012) Annotating simplices with a homology basis and its applications. In: Scandinavian workshop on algorithm theory. Springer, Berlin, pp 189-200
53. Edelsbrunner H, Parsa S (2014) On the computational complexity of Betti numbers: reductions from matrix rank. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 152-160

54. Boissonnat J-D, Dey TK, Maria C (2015) The compressed annotation matrix: an efficient data structure for computing persistent cohomology. Algorithmica 73(3):607-619
55. Mischaikow K, Nanda V (2013) Morse theory for filtrations and efficient computation of persistent homology. Discrete Comput Geom 50(2):330-353
56. Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer P, Vaccarino F (2014) Homological scaffolds of brain functional networks. J R Soc Interface 11(101):20140873
57. Bubenik P (2015) Statistical topological data analysis using persistence landscapes. J Mach Learn Res 16(1):77-102
58. Reininghaus J, Huber S, Bauer U, Kwitt R (2015) A stable multi-scale kernel for topological machine learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4741-4748
59. Kwitt R, Huber S, Niethammer M, Lin W, Bauer U (2015) Statistical topological data analysis - a kernel perspective. In: Advances in neural information processing systems, pp 3070-3078
60. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA (2015) A roadmap for the computation of persistent homology. Preprint. arXiv:1506.08903
61. Fasy BT, Kim J, Lecci F, Maria C, Rouvreau V (2014) TDA: statistical tools for topological data analysis
62. Pearson P, Muellner D, Singh G (2015) TDAmapper: analyze high-dimensional data using discrete Morse theory