Topic 8: Expected Complexity of Topological Summaries

Mihyun Kang and Michael Kerber

January 8, 2024

Scientific background Topological data analysis (TDA) is a novel approach to gain insights on complex data sets using tool from algebraic topology [3]. Since its advent around 25 years ago, the range of application scenarios is ever increasing and nowadays comprises hundreds of papers [6]. One reason of the success of TDA is a well-justified theory that yields simple and explainable summaries on the data sets. The major idea is to split a data set in a collection of "topological features" and a range of scales on which every feature is active. In mathematical terms, this translates into a direct sum decomposition of a graded module obtained by applying homology. In the standard theory, all summands are intervals, resulting in the *persistent barcode* that summarizes the data. Moreover, there are efficient methods to compute such decompositions.

An often observed phenomenon is the discrepancy between the worst-case complexity versus the empirically measured complexity on real-world instances, both for structural and algorithmic properties. For instance, while the barcode computation needs cubic time in the worst case, algorithms for this task show a close-to-linear behavior on practical data sets [10]. Also, in the extension of multi-parameter persistent homology, while arbitrarily complicated summands can occur in the decomposition, empirical measurements rather show that most data set admit a more tame decomposition [2].

A natural approach to shed light on such discrepancies is via the expected instead of the worst-case complexity. Topological properties of random structures are well-studied by now, pioneered by the work of Linial and Meshulam [9] and Kahle [8, 7]. These are extensions of connectivity questions of random graphs, including the famous phase-transition results by Erdős and Rényi. Such results have recently been used to show that the expected complexity of barcode computation is better than what the worst case predicts [5]. It was also shown recently that under fairly general assumptions, a quarter of all summands in a multi-parameter decomposition are simple [1].

Hypotheses/Aims The aforementioned discrepancy between worst-case bounds and empirical measurements in topological data analysis leads to the hypothesis that worst-case instances are rare and artificial (indeed, known worst-case constructions require a careful design to become "bad", e.g. [5, Sec.7]). We further hypothesize that with *random models* that share characteristics of real-world data sets, we can prove that the worst case differs from the *average case*, providing provable evidence for the first hypothesis. Initial results for this hypothesis have been published recently, and our goal is a much deeper investigation of this phenomenon. This requires, in turn, a more extensive study of topological and geometric properties of random graphs and simplicial complexes.

Approaches/Methods We have several direct follow-up questions in accordance to the project goal: for instance, the expected complexity result from [5] is currently limited to 1-dimensional homology because it uses results on the expected homology of random complexes which are only proved in dimension 1 [4, 7]. An extension to higher dimensions would immediately generalize the result and would be of independent interest. Also, while we proved that at least a quarter of all summands in multi-parameter persistence are intervals (the simplest possible form), we observed experimentally that typically, not all summands are intervals [1]. We have also identified a local geometric obstruction to get an interval, and we conjecture that this implies that the probability of only having intervals reaches 0 in the limit.

This research area is a fruitful interplay between experimental and theoretical approaches: both initial results stem from initial experimental evaluations which led us to conjectures that we were able to prove later on. We foresee that this approach will lead to further conjectures and theorems. For instance, can we identify and theoretically analyze a non-trivial multi-parameter random model for which the probability of decomposing in intervals is large? This would be highly interesting because there are numerous algorithms to treat modules with a simple decomposition structure more efficiently.

Time frame The PhD student in this topic must both acquire substantial knowledge in topological data analysis and in the theory and techniques of random graphs and complexes. At the same time they must have excellent mathematical skills but also the ability to do experimental evaluations. This is a long list of requirements and we cannot expect that any PhD candidate brings a background in all these fields. We therefore foresee an intensive training period of 12-18 months to bring the student in the position to start working on the project goal.

Participating faculty members This project is supervised by Mihyun Kang and Michael Kerber.

Mihyun Kang is a full professor at TU Graz and leads the Combinatorics Group. Her main research areas are combinatorics, discrete probability, and algorithms. She is best known for her work on topological properties of random graphs. She received a prestigious *Friedrich Wilhelm Bessel Research Award* of the Alexander von Humboldt Foundation. She serves on the editorial board of leading journals, including *Random Structures and Algorithms*. She supervised five PhD students and currently supervises two postdocs and two PhD students.

Michael Kerber is a full professor at TU Graz and an internationally recognized expert in computational topology and geometry. He works mostly on fast algorithms for computing and comparing persistence diagrams, and on multi-parameter persistent homology. Kerber was program committee chair of the *2022 Symposium on Computational Geometry*, the main conference in computational geometry and topology. Four PhD students have completed their thesis under his supervision. His group currently consists of one postdoc and 4 PhD students.

References

[1] A. Alonso and M. Kerber. Decomposition of zero-dimensional persistence modules via rooted subsets. In *39th International Symposium on Computational Geometry (SoCG 2023)*, 2023.

[2] U. Bauer, M. Botnan, S. Oppermann, and J. Steen. Cotorsion torsion triples and the representation theory of filtered hierarchical clustering. *Advances in Mathematics*, 369:107171, 2020.

[3] G. Carlsson. Topology and data. *Bulletin of the AMS*, 46:255–308, 2009.

[4] B. Demarco, A. Hamm, and J. Kahn. On the triangle space of a random graph. *Journal of Combinatorics*, 4(2):229–249, 2013.

[5] B. Giunti, G. Houry, and M. Kerber. Average complexity of matrix reduction for clique filtrations. In *Proceedings of the 2022 International Symposium on Symbolic and Algebraic Computation*, pages 187–196, 2022.

[6] B. Giunti, J. Lazovskis, and B. Rieck. DONUT: Database of original & non-theoretical uses of topology, 2022. `https://donut.topology.rocks`.

[7] M. Kahle. Random geometric complexes. *Disc. & Comp. Geometry*, 45(3):553–573, 2011.

[8] M. Kahle. Topology of random simplicial complexes: a survey. In *Algebraic topology: applications and new directions*, volume 620 of *Contemp. Math.*, pages 201–221. 2014.

[9] N. Linial and R. Meshulam. Homological connectivity of random 2-complexes. *Combinatorica*, 26:475–487, 2006.

[10] N. Otter, M. Porter, U. Tillmann, P. Grindrod, and H. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.