Towards Testing the Significance of Branching Points and Cycles in Mapper Graphs

Patrik Zajec
patrik.zajec@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Primož Škraba
p.skraba@qmul.ac.uk
School of Mathematical Sciences, Queen Mary University of London
London, UK

Dunja Mladenič
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

Given a point cloud P, which is a set of points embedded in $\mathbb{R}^d$, we are interested in recovering its topological structure. Such a structure can be summarized in the form of a graph. An example of this is the mapper graph, which captures how the point cloud is connected and reflects the branching and cyclic structure of P as branching points (vertices with degree greater than 2) and cycles in the graph. However, such a representation is not always accurate, i.e., the structure shown by the graph may not be sufficiently supported in the point cloud. To this end, we propose an approach that uses persistent (relative) homology to detect branching and cyclic structure, and employs a statistical test to confirm whether the structure is indeed significant. We show how the approach works for low-dimensional point clouds, and discuss its possible applications to real world point clouds.

KEYWORDS

topological data analysis, statistical hypothesis testing, persistent homology, mapper algorithm

1 INTRODUCTION

Consider the point cloud P consisting of points in $\mathbb{R}^2$ shown in Figure 1a. Using the mapper algorithm, we can construct a graph that represents its topological structure like the one in Figure 1b, which seems to recover the important structure. Using the same algorithm (but with different values of its adjustable parameters) we could end up with different graphs. The second graph, shown in Figure 1c, contains two cycles: the middle one, which captures the cycle present in P, and the top one, where the algorithm "mistakenly" considers the top points to connect in a cycle. The third graph, shown in Figure 1d, shows a similar structure as the graph in Figure 1b, although it contains one branching point more (splitting off the upper left branch) and a cycle of length three. One could argue that these branching and cyclic structures are not sufficiently supported in P.

Our goal is to develop an approach that allows us to confirm, through a statistical test, whether the structure recovered by the mapper graph is indeed present in the point cloud. We use persistent homology, a well-known construction from topological data analysis (TDA), to represent the structure from the point cloud, and a recently introduced hypothesis testing framework [1] that provides a way to evaluate the significance of such a structure. We demonstrate the approach on two examples: a Y-shaped point cloud and a sample of a 3D mesh resembling an ant. These low-dimensional examples allow us to visually inspect the results, laying the groundwork for extensive experiments with higher-dimensional point cloud data used in real-world applications.

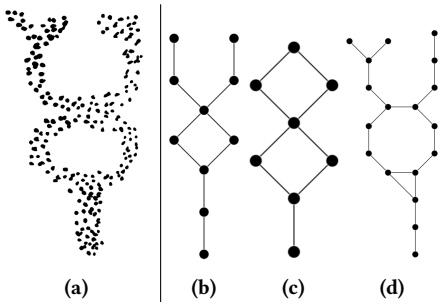

Figure 1: A point cloud (a) and three graphs (b, c, d) summarizing its topological structure, constructed by the mapper algorithm for different choices of its parameters.

Representing the topological structure of the point cloud with a simpler object, such as a graph, and having a statistical method for testing the significance of such a structure is a very relevant task. A simpler representation allows us to visualize [3] and interpret high-dimensional representations that are everywhere in modern data science and machine learning. It might even allow us to find singularities that often carry relevant information. The mapper algorithm [6] is a commonly used tool in TDA. Although it is simple, the result is sensitive to the choice of its parameters [2]. Nevertheless, it provides only one possible low-dimensional view of the input data, and to our knowledge there is no method that would confirm the significance of the represented structure. There is another method, called persistent homology, which, while not directly applicable to visualization, deals with a particular structure of "holes" in space and now has a framework [1] that allows us to statistically test the significance of such a structure.

2 BACKGROUND

A point cloud P is a set of points embedded in $\mathbb{R}^d$ which can be viewed as a sample of a topological space $\mathbb{X}$. Since discrete points from P have no interesting topological structure, we consider the space $P^r = \bigcup_{p \in P} B(p, r)$ for some radius r. If P is a sufficiently dense sample of $\mathbb{X}$, then P^r has some of the same properties as $\mathbb{X}$ for a suitable r. To compute the properties of interest, we represent P^r with a simplicial complex K which, if properly constructed, has homology groups isomorphic to those of P^r. We

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

are interested in finding the branching and cyclic structure in the point cloud, both of which can be detected using (persistent) homology.

2.1 Simplicial complexes

A *(geometric) simplicial complex* K can be thought of as a "high-dimensional graph" whose vertices are points from the point cloud and connectivity is determined by the geometric configuration of the points. In addition to vertices and edges, we include triangles, tetrahedra and higher dimensional simplices. Formally, K consists of finite nonempty subsets of P and is closed under inclusion (i.e., $A \in K$ and $B \subset A$ implies $B \in K$). We refer to elements in K of size $k+1$ as k-simplices, which correspond to k-cliques when we think about K as a hyper-graph.

The Čech and Vietoris-Rips complexes are the two most common constructions, both parameterized by a scale parameter (radius) $r > 0$. We use the Vietoris-Rips construction, where we include a subset of (k + 1) points from P as a k-simplex if all points are at most r apart.

We can construct a sequence of complexes $K_{r_1}, K_{r_2}, \ldots$ by increasing the radius r. Such a construction is "increasing" in the sense that for $r_1 < r_2$, it holds that $K_{r_1} \subseteq K_{r_2}$. Such sequences are also known as *filtrations* and are used in persistent homology.

2.2 Persistent relative homology

Homology. *Homology* is a classical construction in algebraic topology that deals with topological properties of a space. More precisely, it provides a mathematical language for the holes in a topological space. Homology groups denoted by $H_k(\mathbb{X})$, where k is a dimension, capture the holes indirectly by focusing on what surrounds them. For example, the basis of $H_0(\mathbb{X})$ corresponds to the connected components and the basis of $H_1(\mathbb{X})$ to the closed loops surrounding the holes. The rank of the k-th homology group, also known as *Betti number*, counts the number of k-dimensional "holes".

We can construct homology groups for a given simplicial complex K. The important concepts in the construction are: (i) the chain groups C_k, where the k-th chain group consists of all formal linear combinations of k-dimensional simplices $\sum_i a_i \sigma_i$, where σ_i are k-simplices from K and a_i are coefficients, usually from $\mathbb{Z}_2$, (ii) the boundary operator ∂_k, which is a map describing how (k - 1)-simplices are attached to k-simplices, (iii) the groups Z_k of k-cycles, which are k-chains in the kernel of ∂_k, and (iv) the groups B_k of k-boundaries, which are elements in the image of ∂_{k+1}. The boundary operator ∂_k has the property that $\partial_k \circ \partial_{k+1} = 0$, i.e., it maps the boundary of the boundary to zero. Therefore, $B_k \subseteq Z_k$.

Intuitively, a k-cycle can be thought of as a generalized version of a cycle in a graph - it is a sequence of k-dimensional simplices wrapped around something. If this sequence is actually a boundary of a (k+1)-dimensional chain, then its interior is full (trivial cycle). Otherwise, it surrounds a hole. The k-th homology $H_k = ker\ \partial_k / im\ \partial_{k+1} = Z_k / B_k$ takes a "modulo" of k-cycles with k-boundaries, leaving only cycles that are nontrivial.

Relative homology. Given a simplicial complex K and a subcomplex $L \subseteq K$, the relative homology of a pair of topological spaces (simplicial complexes in our case) can be thought of as the (reduced) homology of the quotient space K/L. Intuitively, we want to factor out L, which is expressed by the quotient operation $C_k(K, L) = C_k(K)/C_k(L)$. The group of k-cycles becomes $Z_k(K, L) = Z_k(K)/Z_k(L)$, which we call the group of *relative cycles*. We can think of the reduced homology of a space as if we were representing the entire L with a single point.

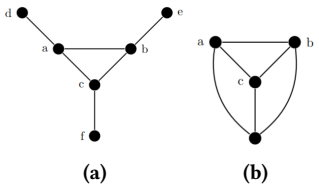

Figure 2: a) A Y-shaped simplicial complex with one cycle. b) The quotient K/L, where subcomplex L contains 0-simplices {d, e, f}. Such identification introduces two new 1-dimensional "holes", captured by the relative homology group $H_1(K, L)$.

The concept of homology and relative homology is best illustrated by an example. Consider a simple simplicial complex consisting of 0-simplices {a, b, c, d, e, f} and 1-simplices {(a, b), (a, c), (a, d), (b, e), (c, f)} as shown in Figure 2a. There is a "hole" of dimension 1 (surrounded by the cycle $a \to b \to c \to a$), which is captured in the homology group H_1. Choosing $L = \{d, e, f\}$ as a subcomplex, the quotient K/L identifies the simplices from L to a single point, as shown in the figure 2b. This results in two new "holes" in dimension 1, which are captured by the relative homology group $H_1(K, L)$, which has rank 3. This "lifting property" of relative homology (introducing new "holes" when identifying simplices) is used in our approach to detect branching points.

Persistent homology. The constriction of the simplicial complex and hence the groups H_k are highly sensitive to the choice of radius r. To overcome this, persistent homology considers the entire range of scales and tracks the evolution of k-cycles as the value of r increases, thus forming a sequence of filtrations. In this process, cycles are created (born) and later filled-in (die). This information is most often represented by *persistence diagrams*, a two dimensional scatter plot, $dgm_k = \{p1, \ldots, p_m\}$, where each point $p_i = (b_i, d_i)$ represents the birth and death times (radius) of the associated persistent cycle.

2.3 Significance testing of persistent cycles

The significance of topological features is often measured by the lifetimes of persistent cycles, i.e., $\delta = (d_i - b_i)$. Although this method is intuitive as it captures the geometric “size” of topological features, [1] uses the statistic $\pi_i = d_i/b_i$. They present a statistical test to determine for each point $p_i \in dgm_k$ whether it is a signal or noise, i.e., a significant structure or the result of noise and randomness in the data. They introduce a special transformation $l(p_i)$ applied to each point from the diagram where the values of $l(p_i)$ follow a certain (LGumbel) distribution if p_i are points corresponding to noisy cycles, while cycles significantly deviating from this distribution are declared as signal. The signal part of dgm_k can be recovered as $dgm_k^s(\alpha) = \{p \in dgm_k : e^{-e^{l(p)}} < \frac{\alpha}{|dgm_k|}\}$ given a p-value α.

Computing persistent homology for an entire filtration is often intractable, as higher values of r lead to a large number of simplices. The common practice is to set a threshold r_{max} and calculate $dgm_k(r_{max})$ using simplices generated up to r_{max}. This often leads to cycles that are "infinite", i.e., born prior to r_{max} but die after r_{max}. The framework also provides an algorithm to

determine the infinite cycles that are already significant, and provides means to select the next r_{max} threshold to inspect infinite cycles that have not yet been determined to be significant.

2.4 The mapper algorithm

Given the topological space $\mathbb{X}$ and a continuous function $f : \mathbb{X} \to \mathbb{R}$, the mapper algorithm [6] constructs a graph $G = (V, E)$ that captures the topological structure of $\mathbb{X}$. It does so by pulling back a cover $\mathcal{U}$ of the space $f(\mathbb{X})$ to a cover on $\mathbb{X}$ through f. We can view the function f and the cover $\mathcal{U}$ as the lens through which the input data $\mathbb{X}$ is examined.

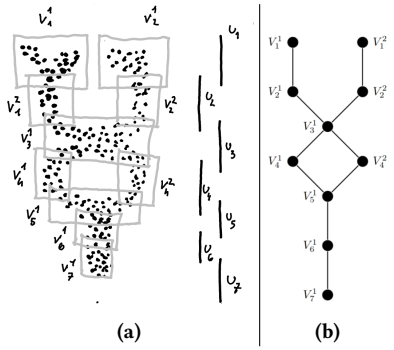

Figure 3: An example of the construction of a mapper graph. (a) A 2-dimensional point cloud P with cover $\{V_i^j\}$, a function $f : \mathbb{R}^2 \to \mathbb{R}$ and cover $\mathcal{U}$ of $f(P)$. (b) The resulting mapper graph.

Given a point cloud P and $f : P \to \mathbb{R}$, we first construct a set of n intervals $\mathcal{U} = \{U_1, \ldots, U_n\}$ covering $f(P)$. The percentage of overlap for two consecutive intervals U_i and U_{i+1} is determined by the parameter p. For each interval $U_i = (a, b)$, let $P_{U_i} = f^{-1}(U_i)$ be a set of points with function values in the range (a, b). The set P_{U_i} for each U_i is further partitioned into $V^1, \ldots, V^{k_i}$ by a clustering algorithm (in our case DBSCAN [5] with parameter ϵ, which sets the maximum distance between two samples so that one is considered to be in the neighborhood of the other) to obtain a cover of $P = \bigcup_{i=1,\ldots,n}\{V_i^1, \ldots, V_i^{k_i}\}$. Each $V_i^j \subset P$ becomes some vertex v in the mapper graph with $\phi(v) = V_i^j$ mapping v to a subset of points. Two vertices are connected by an edge if their point sets intersect (see Figure 3).

The resulting graph $G = (V, E)$ provides a combinatorial description of the data and the mapping $\phi : V \to \mathcal{P}(P)$ maps each node $v \in V$ to a subset of points from P.

3 METHODOLOGY

The input to our approach is a set of points P embedded in $\mathbb{R}^d$ and a graph $G = (V, E)$ together with a mapping $\phi : V \to \mathcal{P}(P)$ that maps each vertex to a subset of points. Note that the method used to construct the graph is not limited to the mapper algorithm.

The graph is assumed to capture the topological structure of the point cloud, i.e., branching points (vertices with a degree of at least 3) and cycles in the graph should reflect the branching and cyclic structure of the point cloud. Our approach tests whether the captured structure is significant when viewed through homology, operating directly on a subset of points from the point cloud.

3.1 Testing the cycles

A *simple cycle* is a finite sequence of vertices $v_1 \to v_2 \to \ldots \to v_n$, where v_i and v_{i+1} are connected by an edge such that no vertex, except the endpoint, repeats ($v_i = v_j$ if and only if $i, j \in \{1, n\}$). Let $v_1, \ldots, v_n$ be such a cycle from G. We compute the persistence diagram of the subset $P' = \bigcup_{i=1,\ldots,n} \phi(v_i)$ and use the test [1] to confirm that it contains at least one significant cycle ("hole") of dimension 1.

3.2 Testing the branching structure

Let $N(v)$ be a set of vertices connected to v (1-hop neighborhood) and let v be a branching point in G (as in Figure 4). Let $N'(v) = \{u : u \in N(v), deg(u) \geq 2\}$ be a set of vertices from $N(v)$ that have at least one additional neighbor. Together with v, $N'(v)$ forms a set of internal points $I_v = \bigcup_{u \in \{v\} \cup N'(v)} \phi(u)$ (shown in Figure 4 as black vertices inside the outer black line).

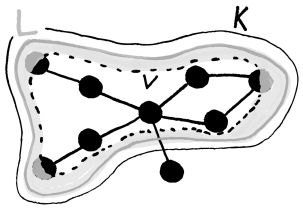

Figure 4: Construction of K and L for a branching point v. Vertices forming K are inside the outer black line. Vertices forming L are bicolored, indicating that some of their points are inside due to overlap between the vertices' point sets.

Let $K_v = \bigcup_{u \in N'(v)} N(u)$ be a set of vertices whose points are used to form a complex K (vertices inside the outer black line in Figure 4), i.e. K is formed from the points $\bigcup_{u \in K_v} \phi(u)$. Now let L be a subcomplex of K containing simplices which do not contain any of the points from I_v. Thus L contains points of vertices exactly two edges away from v (bicolored vertices in Figure 4). We use K and L to compute relative persistent homology, identifying simplices of L to a single point and introducing relative cycles ("holes") when $K \backslash L$ has a branching structure. For a branching point v, the relative persistence diagram should contain at least $deg(v) - 1$ significant relative cycles.

4 EXPERIMENTS

We perform experiments illustrating our approach on two point clouds. The graphs are constructed using the mapper algorithm from the Giotto TDA library [7] with the parameters specified for each experiment. To construct the simplicial complex and compute (relative) persistent homology, we use the Dionysus library [1]. We increase the initial radius r using the algorithm from [1] until either no infinite cycles remain or all currently infinite cycles are identified as significant.

We include a figure of the graph for each experiment and mark interesting branching points and cycles. The points corresponding to a cycle are shown in red, the internal points of a branching point are also red, while the boundary points (forming L) are blue.

[1] Available at: https://github.com/mrzv/dionysus.

4.1 Experiment 1: Y-shaped point cloud

The point cloud P consists of 5000 points in $\mathbb{R}^2$ and resembles a Y-shape with a cycle in the centre. The graph (see Figure 5) was created with the following parameters: f is a projection on the x-coordinate, $n = 30$, $p = 0.5$ and $\epsilon = 3$.

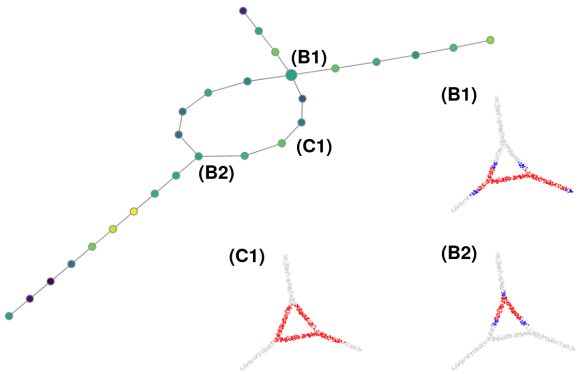

Figure 5: Mapper graph with two branching points (B1 and B2) and one simple cycle (C1) together with their corresponding subsets of points.

The graph contains one simple cycle, which is also significant because the subset of its points contains a homologically significant cycle. The graph also contains two branching points, B1 and B2 with degrees 4 and 3.

The persistence diagram for B1 has three (significant) infinite cycles, indicating a branching structure of degree 4, while the diagram for B2 has two (significant) infinite cycles, indicating a branching structure of degree 3. In this example, it was confirmed that both the cyclic and the branching structure of the graph are reflected in the point cloud.

4.2 Experiment 2: 3D ant surface

The point cloud P consists of 6370 points in $\mathbb{R}^3$ corresponding to the vertices of a 3D mesh in the form of an ant obtained from [4]. The graph (see Figure 6) was created with the following parameters: f is the distance to the tip of the ant's abdomen, $n = 50$, $p = 0.5$, and $\epsilon = 0.025$.

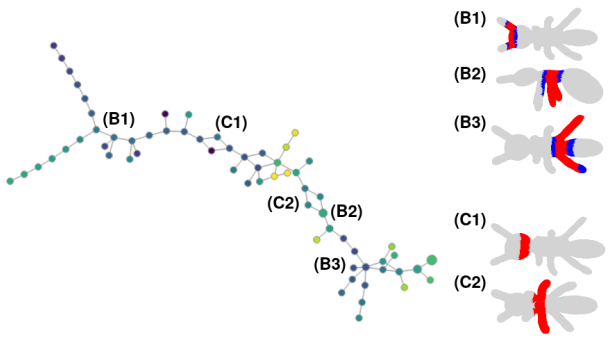

Figure 6: Mapper graph with three highlighted branching points (B1, B2 and B3) and two simple cycle (C1, C2) together with their corresponding subsets of points.

We highlight three interesting branching points. Vertex B1 is a branching point of degree 3, which corresponds to the branching on the ant's head into its two antennas and is significant. Vertex B2 is a branching point of degree 3 and one of the vertices from the cycle C2. Looking at the point cloud, no branching structure is detected because the points of the two legs are contained in the vertex B2 itself and there are no boundary points on the legs, so they appear as a single connected blob. Our approach does not detect a branching structure, even though there is, as some other strategy of selecting the boundary points would need to be used. Vertex B3 has degree 6, but only 5 neighbors are used as one does not have any additional neighbor except B3. Since one of the legs has no boundary points, only 2 cycles appear, causing B3 to be recognized as a branching point with degree 3.

We also highlight 2 simple cycles. Cycle C1 wraps around the ant's hollow head and is recognized as significant. Cycle C2 wraps around the ant's two middle legs and part of its body. No significant cycles were found - ant's legs are not close enough together to form a large cycle and cycle formed by the hollow legs is too small to be detected. So there is not enough support to confirm the structure found by mapper.

5 CONCLUSIONS AND FUTURE WORK

We have demonstrated, how persistent (relative) homology can be used in conjunction with a statistical test to confirm the significance of the topological structure of a point cloud summarized with a graph. In the future, we will conduct extensive experiments on more complex, high-dimensional point clouds with known and unknown structure. Ideally, we could use our approach to prune the mapper graphs or guide the selection of values for its parameters. Our approach to identifying branching structures needs further work, as the current strategy of using a (modified) 2-hop neighborhood as a boundary sometimes fails. In addition, we may need a more sensitive version of the statistical test from [1] which is currently stated to hold in general but might be possible to adapt for a particular type of data.

ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency under the project J2-1736 Causalify and co-financed by the Republic of Slovenia and the European Union's HE program under enRichMyData EU project grant agreement number 101070284.

REFERENCES

[1] Omer Bobrowski and Primoz Skraba. 2023. A universal null-distribution for topological data analysis. *Scientific Reports*, 13, 1, (July 2023), 12274. DOI: 10.1038/s41598-023-37842-2.

[2] Mathieu Carrière, Bertrand Michel, and Steve Oudot. 2018. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19, 12, 1–39. http://jmlr.org/papers/v19/17-291.html.

[3] Nithin Chalapathi, Youjia Zhou, and Bei Wang. 2021. Adaptive covers for mapper graphs using information criteria. In *2021 IEEE International Conference on Big Data (Big Data)*, 3789–3800. DOI: 10.1109/BigData52589.2021.9671324.

[4] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. 2009. A benchmark for 3d mesh segmentation. *ACM Trans. Graph.*, 28, 3, Article 73, (July 2009), 12 pages. DOI: 10.1145/1531326.1531379.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In (KDD'96). AAAI Press, Portland, Oregon, 226–231.

[6] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. 2007. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Eurographics Symposium on Point-Based Graphics*. M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors. The Eurographics Association. ISBN: 978-3-905673-51-7. DOI: 10.2312/SPBG/SPBG07/091-100.

[7] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. 2020. Giotto-tda: a topological data analysis toolkit for machine learning and data exploration. (2020). arXiv: 2004.02551 [cs.LG].