Pattern Recognition Letters 147 (2021) 85–92

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Stable topological signatures for metric trees through graph approximations[☆]

Robin Vandaele[a,b,c,*], Bastian Rieck[d,e], Yvan Saeys[b,c], Tijl De Bie[a]

[a] IDLab, Department of Electronics and Information Systems, Ghent University, Gent, Belgium
[b] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium
[c] Data mining and Modelling for Biomedicine (DaMBi), VIB Inflammation Research Center, Gent, Belgium.
[d] Machine Learning and Computational Biology Lab, D-BSSE, ETH Zurich, Switzerland
[e] SIB Swiss Institute of Bionformatics, Switzerland

ARTICLE INFO

Article history:
Received 10 May 2020
Revised 14 January 2021
Accepted 26 March 2021
Available online 14 April 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Topological data analysis
Algebraic topology
Persistent homology
Proximity graphs
Metric trees
Cell trajectory inference

ABSTRACT

The rising field of *Topological Data Analysis* (TDA) provides a new approach to learning from data through *persistence diagrams*, which are *topological signatures* that quantify topological properties of data in a comparable manner. For point clouds, these diagrams are often derived from the Vietoris-Rips filtration—based on the metric equipped on the data—which allows one to deduce topological patterns such as components and cycles of the underlying space. In *metric trees* these diagrams often fail to capture other crucial topological properties, such as the present leaves and multifurcations. Prior methods and results for persistent homology attempting to overcome this issue mainly target Rips graphs, which are often unfavorable in case of non-uniform density across our point cloud. We therefore introduce a new theoretical foundation for learning a wider variety of topological patterns through *any given graph*. Given particular powerful functions defining persistence diagrams to summarize topological patterns, including the *normalized centrality* or *eccentricity*, we prove a new stability result, explicitly bounding the *bottleneck distance* between the true and empirical diagrams for metric trees. This bound is tight if the metric distortion obtained through the graph and its maximal edge-weight are small. Through a case study of gene expression data, we demonstrate that our newly introduced diagrams provide novel quality measures and insights into cell trajectory inference.

1. Introduction

For the past decade, *persistent homology* [16]—the most prominently used and studied tool within the field of *Topological Data Analysis* (TDA) [6]—has led to many new applications to supervised and unsupervised machine learning. Many of the data sets to which persistent homology has been successfully applied, were already at least partially structured, in the form of a *simplicial complex*, i.e., a higher-dimensional generalization of a graph. Examples of these include brain networks [17], meshes [14], and images [2,27]. Persistent homology then tracks topological changes over a *filtration*, i.e., a nested sequence of subcomplexes of the original complex.

In the case of point cloud data, the data is often sampled from a topological structure, the knowledge of which provides tremendous insight into the underlying structure or data generating process. However, the underlying topology is often difficult to reveal, due to the high dimensionality of the data, or noise. Since they lack a naturally induced simplicial structure, computing persistent homology of point clouds is mostly feasible through the the *Vietoris-Rips filtration* [22]. Unfortunately, this type of *persistence*—a measure of prominence or relevance of a topological feature—is often insufficient, as it merely detects gaps, cycles, voids, and higher-dimensional holes in the model. Thus, it is impossible to distinguish between point clouds sampled from a linear ('I'-shaped) versus a bifurcating ('Y'-shaped) topology through this method.

We therefore develop a new foundation for learning topological patterns through *graph approximations*. These graphs will be used as simplicial representation of the data. As will be shown in this paper, they allow us to learn a wider variety of topological patterns in *metric trees*, in theory and in practice.

[☆] Editor: Nicolas Passat

[*] Corresponding author at: IDLab, Department of Electronics and Information Systems, Ghent University, Gent, Belgium.

E-mail address: Robin.Vandaele@UGent.be (R. Vandaele).

https://doi.org/10.1016/j.patrec.2021.03.035

1.1. Contributions

- We provide an intuitive introduction to, as well as a formal theoretical foundation for studying topological patterns through 0-dimensional persistence of arbitrary graph approximations (Section 2).
- We show under which conditions functions lead to a non-trivial *stability* result—guaranteeing that our true and empirical persistence diagram are close—for graph approximations (Theorems 2.1 and 2.2). We provide two such functions quantifying powerful topological features of metric trees: the *eccentricity* and *normalized centrality* (Corollary 2.1).
- We introduce a novel application of our signatures that goes beyond standard topological inference, providing novel quality measures and insights to the field of cell trajectory inference (Section 3.2).
- We summarize how our method leads to and opens up new possibilities for learning topological patterns (Section 4).

1.2. Background on persistent homology

The concept of *persistent homology* has its roots in the field of *algebraic topology* [18]. Its computation requires two things: a *simplicial complex* K, and a *filtration* $\mathcal{F}$ defined on K. A simplicial complex can be seen as a generalization of a graph, that apart from 0-simplices (nodes) and 1-simplices (edges), may also include 2-simplices (triangles), 3-simplices (tetrahedra), and so on. A simplicial complex K is furthermore closed under inclusion, i.e., if $\sigma' \subseteq \sigma \in K$ then $\sigma' \in K$. A filtration $\mathcal{F}$ on K is then a nested sequence $K_0 \subseteq K_1 \subseteq \ldots \subseteq K_N = K$ of subcomplexes of K. Fig. 2a illustrates these concepts by means of a point cloud data set D sampled from the unit circle. Here, the filtration equals the *Vietoris-Rips filtration* $\mathrm{VR}^k(D)$, defined as the nested sequence

$$\left(\mathrm{VR}^k_\epsilon(D) := \{S \subset D : |S| \leq k+1 \wedge \operatorname{diam}(S) \leq \epsilon\}\right)_\epsilon,$$

parameterized by the *time* $\epsilon \in \mathbb{R}_{\geq 0}$. The *Vietoris-Rips complex* $\mathrm{VR}^k_\epsilon(D)$ contains all simplices in D of diameter less than or equal to ϵ, and of *dimension* less than or equal to k. If $k = 1$, we simply refer to the complex as the *(Vietoris-) Rips graph*.

The number of k-dimensional holes in a complex is expressed by the Betti number β_k. In this sense, a 0-dimensional hole is a ‘gap’, and β_0 corresponds to the number of connected components, β_1 corresponds to the number of loops, β_2 to the number of voids, and so on. Persistent homology quantifies topological changes through the *birth* and *death* of these holes across the filtration. E.g., in Fig. 2a, every data point corresponds to the birth of a connected component at the start of the filtration. By increasing ϵ, points get connected to each other, resulting in the death of many of these components. From around $\epsilon = 0.75$, the complex consists of one connected component, as well as a loop representing the underlying cyclic structure.[1] Increasing ϵ further, this loop gets ‘filled in’ through the 2-simplices, resulting in its death. The idea behind persistent homology and persistence is that holes persisting for a long range of consecutive values ϵ represent significant features of the topology underlying the point cloud. This is illustrated by the *persistence diagrams* $\mathcal{D}_k$ (one for each considered dimension $k \in \{0, 1\}$ of holes) in Fig. 2b. This is a multiset containing a point (b, d) for each hole that was born at $\epsilon = b$ and died at $\epsilon = d$. By definition, $d = \infty$ if a hole never dies. These points are usually displayed at the top of the diagram. Furthermore, by convention, a persistence diagram contains every point on the diagonal.

To understand one of the most important concepts in TDA (and in this paper), i.e., *stability*, we first need to introduce some definitions [1,22].

Definition 1.1. Let $\mathcal{D}$ and $\mathcal{D}'$ be two persistence diagrams. The *bottleneck distance* between them is defined as

$$d_\mathrm{b}\left(\mathcal{D}, \mathcal{D}'\right) := \inf_\varphi \sup_x \|x - \varphi(x)\|_\infty \in \mathbb{R}_{\geq 0} \cup \{\infty\},$$

where φ ranges over all bijections from $\mathcal{D}$ to $\mathcal{D}'$, and x ranges over all points in $\mathcal{D}$. Since the diagrams include the diagonal, $|\mathcal{D}| = |\mathcal{D}'| = |\mathbb{R}|$. Thus, $d_\mathrm{b}\left(\mathcal{D}, \mathcal{D}'\right)$ is well-defined.

Definition 1.2. Let (X, d_X) and (Y, d_Y) be two metric spaces. A *correspondence* is a set $C \subseteq X \times Y$, such that for any $x \in X$, there exists $y \in Y$ such that $(x, y) \in C$, and vice versa. Given $\epsilon \in \mathbb{R}^+$, a correspondence C is an *ϵ-correspondence* if $(x, y), (x', y') \in C$ implies that $|d_X(x, x') - d_Y(y, y')| \leq \epsilon$. The *Gromov-Hausdorff distance* $d_{\mathrm{GH}}(X, Y)$ is the infimum of the ϵ for which there exists an ϵ-correspondence between (X, d_X) and (Y, d_Y).

Stability ensures that if two finite metric spaces are close, their persistence diagrams obtained through the Vietoris-Rips filtrations are close as well. More formally, if (X, d_X) and (Y, d_Y) are two finite metric spaces, then [11]

$$d_\mathrm{b}(\mathrm{Dgm}_k(\mathrm{VR}(X)), \mathrm{Dgm}_k(\mathrm{VR}(Y))) \leq 2d_{\mathrm{GH}}(X, Y).$$

Stability results formulated through the ground truth topology also exist for the Vietoris-Rips filtration, but their formulation tends to be more complicated [22].

1.3. Related work

Persistent homology has already been used extensively in (un)supervised machine learning problems. In this context, it can be regarded as a feature engineering method, where its resulting persistence diagrams correspond to *topological signatures*, encoding structural information at varying scales in the data. Our purpose is not to outperform these methods, but rather to extend them to become applicable to a wider variety of data sets for which learning topological patterns remains an important challenge—in our case—*metric trees*.

The main novelty of our introduced stability result (Theorem 2.2) is its generality in terms of the type of graph approximation, instead of its generality in terms of the dimension of persistent homology. In case of metric trees, we will show that the restriction to 0-dimensional persistence is indeed sufficient for revealing multifurcations and leaves. However, the restriction to particular graphs such as Rips graphs is often unfavorable in case of non-uniform density across our point cloud.

That being said, it is worth pointing out the differences of our work to the following.

TDA through functions

The idea of TDA through functions equipped on point cloud data, and in particular, the *eccentricity function* (Corollary 2.1), is not novel. Indeed, Carlsson [7] previously discussed that (regular) persistent homology through the Vietoris-Rips complex may miss out on finding meaningful structure in many examples of point cloud data. He proposed a refinement under the name of *functional persistence*. The idea is to apply regular persistence to a subset of the data, obtained through thresholding according to a user defined function. E.g., by applying regular persistence to a subset of points sufficiently far away from the center of the point cloud, one may be able to deduce a flare-structured topology. However, his introduction to functional persistence is rather brief, and this method mainly serves as a visual inspection tool for individual data sets.

[1] Though many such loops exist, they are all equivalent, representing the same class in the *homology group* associated to this complex [18].

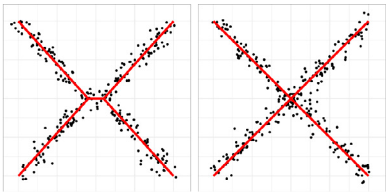
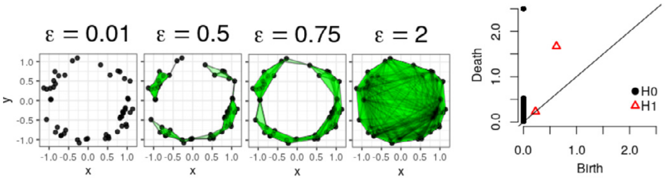


Fig. 1. Point cloud data sets sample from (Left) an H-structured and (Right) an X-structured topology. The ground truth models are shown in red. As the middle branch of the H-structured topology is short relative to the amount of noise in the data, its underlying topology becomes difficult to distinguish from an X-structured topology. The purpose of our current work is to theoretically and practically quantify that these patterns are similar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) An example of four simplicial complexes $\mathrm{VR}^2_\epsilon(D)$.

(b) The diagrams $\mathcal{D}_0$ and $\mathcal{D}_1$ of $\mathrm{VR}^2(D)$.

Fig. 2. Persistent homology through the Vietoris-Rips filtration of a point cloud data set D. The two highly elevated points in the persistence diagram identify the presence of one connected component (H0) and one cycle (H1).

Chazal et al. [10] and Oudot [22] extend persistent homology of metric spaces to metric spaces equipped with a real-valued function f. They vary the Vietoris-Rips complex $\mathrm{VR}^k_\epsilon(D_\epsilon)$ alongside the *sublevel sets* $D_\epsilon := \{x \in D : f(x) \leq \epsilon\}$ on which the simplicial complexes are constructed. Both the simplicial complexes as well as the data on which they are constructed are indexed through the same parameter ϵ. Although their provided stability result applies to persistent homology in any dimension [22, Th. 7.11] it is restricted to Rips based filtrations. Furthermore, this method might only make sense for discovering topological features other than components and cycles whenever the distance metric on D and the functional values f take on a similar scale. In this case, topological features will generally appear less prominent, as higher weight edges will be added at later times (Fig. 3c). Furthermore, inclusion of all vertices would then mean the simplicial complex has to be grown until all pairs of nodes are connected by an edge, making this method computationally less efficient.

Finally, Carrière et al. [9] present a stability result for sublevel filtrations constructed from the ground truth and (a pair of) Rips graphs constructed from a point cloud approximation. In particular Lemma 3.3 by [9] is closely related to our main result (Theorem 2.2), and our restriction to *Lipschitz functions* is inspired by this result. The exact differences between both results will be pointed out in Remark 2.2, after our main theorem.

Persistent local homology

The idea of persistent local homology [3] is to infer topological properties of stratified spaces (including metric trees), by studying persistent homology of the data after removing a neighborhood $B_r(x)$ of a particular point x. This is very similar to the concept of functional persistence, as discussed above. Since it is rather difficult to pinpoint a single suitable radius r, this parameter is often varied as well, resulting in a 1-parameter family of persistence diagrams also known as a *persistence vineyard*. Unfortunately, the current theoretical analysis of this method is again restricted to particular filtrations, such as Rips based filtrations or filtrations based on the *Delaunay triangulation*, the latter of which is challenging to compute in higher dimensional data [4]. Furthermore, existing implementations for computing persistence vineyards are limited (e.g., the Dionysus 1 library in C++), and well-studied methods for comparing persistence vineyards (similar to the bottleneck distance) are lacking.

Mapper

The *Mapper* algorithm mainly serves as a data visualization method, and has been successfully applied to metric trees [21]. The algorithm itself does not directly provide topological signatures, as we do in this paper.

Mapper is quite sensitive to its parameters [19]. Some work to overcome this issue has been performed by Dey et al. [13], under the name of *multiscale Mapper*. The idea is to track the changes in homology of the output of the Mapper algorithm across a varying parameter sequence. However, similar to persistent homology through the Vietoris-Rips filtration, this method only tracks changes in the number of connected components or cycles in the global model.

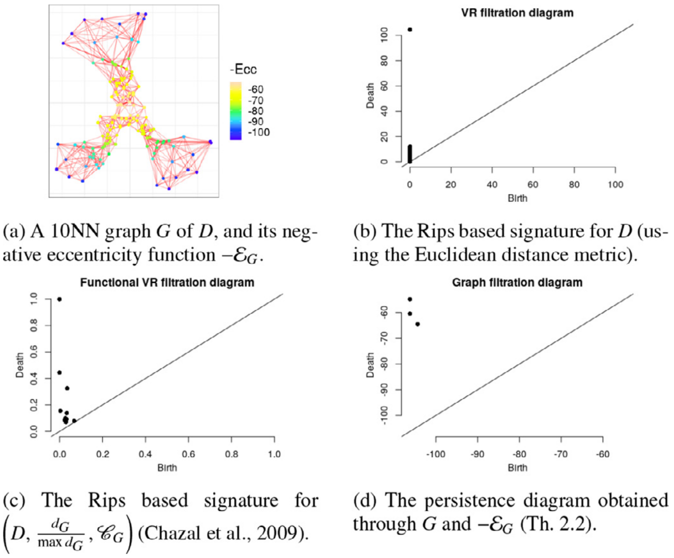
The result of Mapper is commonly a graph. Hence our main result (Theorem 2.2) can also be applied to study how well these graphs preserve topological information of metric trees. In line with this approach, prior results do allow one to quantify the degree of (in)stability of topological features (including leaves) obtained, in case the used clustering method (one is required by the Mapper algorithm) coincides with obtaining connected components in Rips (sub)graphs [8].

Metric graph reconstruction

The case studies in our paper are *graph (tree)-structured topologies*, previously studied by Aanjaneya et al. [1]. This work strongly connects to ours on a theoretical level, as we also formally define the metric distortion we obtain through our graph approximation through the concept of ϵ-correspondence. The major difference is that *we do not require any assumptions on the underlying topology to provide our theoretical guarantee*, which is the bound on the distance between our true and empirical topological signature (Theorem 2.2). By contrast, Aanjaneya et al. [1] require that the metric distortion is bounded by a function of the shortest branch length of the underlying topology to guarantee its reconstruction. For example, one cannot guarantee the correct reconstruction of an H-structured topology if the noise in the data is too high relative to the length of middle branch. In this case, it may become difficult to distinguish the underlying topology from an X-structured topology, as illustrated in Fig. 1.

2. Persistent homology through graph approximations

Fig. 3 b shows that ‘regular’ (0-dimensional) persistent homology of the point cloud data set D shown in Fig. 3a misses out on capturing any topological information other than the underlying model being connected. We can however equip D with a function f that expresses how far a point is from the data center. To this end, we first constructed a 10NN graph G from D, and then computed its negative *eccentricity* function $f = -\mathcal{E}_G$, where $\mathcal{E}_G := \max_{x \in D} d_G(\cdot, x)$. After rescaling both f and the shortest path distance metric d_G on G to [0, 1], the Rips based signature presented by Chazal et al. [10] for the metric space (D, d_G) equipped with resulting *normalized centrality* function $\mathscr{C}_G := \frac{\mathcal{E}_G^{\max} - \mathcal{E}_G(\cdot)}{\mathcal{E}_G^{\max} - \mathcal{E}_G^{\min}}$ now captures some additional structural information. The three ‘leaves’ present in the topology underlying D correspond to the three most elevated points in the diagram (Fig. 3c). However, the components representing these leaves merge quickly before reaching the center of bifurcation, due to the addition of higher weight edges that are not present G. In contrast to this, (0-dimensional) persistent homology of the *sublevel filtration* $(G[\{v \in V(G) : -\mathcal{E}_G(v) \leq t\}])_{t \in \mathbb{R}}$ easily identifies the presence of three leaves. Here, $G[U]$ denotes the subgraph of G induced by the set of nodes $U \subseteq V(G)$.



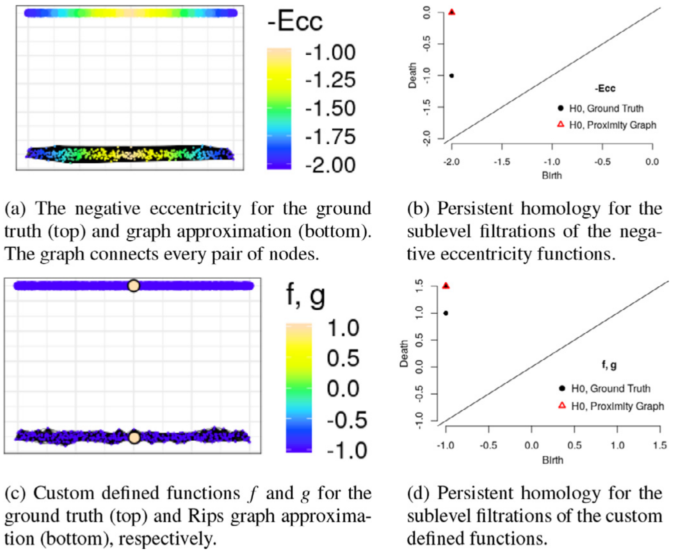
(a) A 10NN graph G of D, and its negative eccentricity function $-\mathcal{E}_G$.

(b) The Rips based signature for D (using the Euclidean distance metric).

(c) The Rips based signature for $\left(D, \frac{d_G}{\max d_G}, \mathscr{E}_G\right)$ (Chazal et al., 2009).

(d) The persistence diagram obtained through G and $-\mathcal{E}_G$ (Th. 2.2).

Fig. 3. (0-dimensional) Rips based signatures for a point cloud data set D, and a custom defined filtration on a 10NN graph G constructed from D. The lower and upper limits of the diagram axes are defined through the first and last 'time' a simplex is added to the complex, respectively.



(a) The negative eccentricity for the ground truth (top) and graph approximation (bottom). The graph connects every pair of nodes.

(b) Persistent homology for the sublevel filtrations of the negative eccentricity functions.

(c) Custom defined functions f and g for the ground truth (top) and Rips graph approximation (bottom), respectively.

(d) Persistent homology for the sublevel filtrations of the custom defined functions.

Fig. 4. In terms of Theorem 2.1, these examples show that a (ϵ-)correspondence can preserve the metrics and function values of f and g arbitrarily well (in terms of ϵ), while simultaneously, $\max\{a, b\}$ can be arbitrarily high.

The purpose of this section is to provide a more formal theoretical foundation for this last type of persistence through graph 'approximations'. The term 'approximations' is to be loosely interpreted, in the sense that we are given some graph that is meant to capture topological information of the data. This can be a Rips graph, kNN graph, minimum spanning tree, or any type of neighborhood graph constructed from the data. Furthermore, this may also be the result of a (graph) model inference method such as the Mapper algorithm.

In Section 2.1, we will illustrate the concept of stability through graph approximations, and discuss the main obstacles for introducing an immediate stability result. In Section 2.2, we prove a new stability result for metric trees.

2.1. Stability through graph approximations

The following theorem states that for any correspondence C between the points in a metric space (X, d_X) and nodes in a graph G, and functions $f : X \to \mathbb{R}$, $g : V(G) \to \mathbb{R}$, one may bound the bottleneck distance between the diagrams for f and g by a value $m = \max\{a, b\}$, measuring how well f and g preserve the connectivity in their respective sublevel filtrations under C.

Theorem 2.1. *Let (X, d_X) be a connected metric space, G a graph, $f : X \to \mathbb{R}$ a tame function, and $g : V(G) \to \mathbb{R}$. Let $a, b > 0$, and suppose $C \subseteq X \times V(G)$ is a correspondence with the following properties:*

- *for all $t \in \mathbb{R}$, if $x \sim y$ in $\{z \in X : f(z) \leq t\}$ and $(x, u), (y, v) \in C$, then $u \sim v$ in $G[w \in V(G) : g(w) \leq t + a]$,*
- *for all $t \in \mathbb{R}$, if $u \sim v$ in $G[w \in V(G) : g(w) \leq t]$ and $(x, u), (y, v) \in C$, then $x \sim y$ in $\{z \in X : f(z) \leq t + b\}$,*

where $\cdot \sim \cdot$ denotes that two points are connected by a path in their respective space (topological or graph), and $G[U]$ denotes the subgraph of G induced by the nodes $U \subseteq V(G)$. Then

$$d_b\left(\mathrm{Dgm}_0\left(\overline{\mathcal{F}_f}(X)\right), \mathrm{Dgm}_0\left(\overline{\mathcal{F}_g}(G)\right)\right) \leq \max\{a, b\},$$

where by $\overline{\mathcal{F}_f}(X)$ (resp. $\overline{\mathcal{F}_g}(G)$), we denote the sublevel filtration $(\{x \in X : f(x) \leq t\})_{t \in \mathbb{R}}$ (resp. $(G[\{v \in V(G) : g(v) \leq t\}])_{t \in \mathbb{R}}$).

Proof. As most definitions in this proof are unimportant for the rest of our paper, they will be omitted for conciseness.

First, observe that $\mathrm{Dgm}_0\left(\overline{\mathcal{F}_g}(G)\right) = \mathrm{Dgm}_0\left(\overline{\mathcal{F}_{|g|}}(|G|)\right)$, where $|G|$ is a *geometric realization* of G and $|g|$ is obtained by extending g on $|G|$ through linear interpolation [15,23]. Now let T_f and $T_{|g|}$ be the *merge trees* of f and $|g|$, respectively [20]. Note that their elements (points) are equivalent classes. Let $\mu := \max\{a, b\}$, and consider the mapping

$$\alpha^\mu : T_f \to T_{|g|} : [(x, t)]_{T_f} \mapsto [(y, t + \mu)]_{T_{|g|}},$$

where y is any node of G such that $(x, y) \in C$. Also consider

$$\beta^\mu : T_{|g|} \to T_f : [(\tilde{y}, t)]_{T_{|g|}} \mapsto [(x, t + \mu)]_{T_f},$$

where $(x, y) \in C$ for some endpoint y of the segment in $|G|$ including $\tilde{y}$, for which $g(y) = |g|(y) \leq |g|(\tilde{y})$. It immediately follows that α^μ and β^μ are *μ-compatible* maps. Furthermore, since by assumption the time increment needed for two points to become connected in one space does not become larger for their corresponding points (under C) after an initial increment by μ, α^μ and β^μ are both continuous and in particular well-defined. The result now follows from [20, Th. 3]. □

Theorem 2.1 cannot yet be interpreted as a stability result. We must still express how the distance between the diagrams depends on the closeness of (X, d_X) and G. However, even if (X, d_X) and G are arbitrarily close in the sense of an ϵ-correspondence C, and $f : X \to \mathbb{R}$ and $g : V(G) \to \mathbb{R}$ are arbitrarily well-preserved under this correspondence, there is generally no guarantee that the diagrams are close as well. This is illustrated by two example models and their graph approximations in Fig. 4.

In the first example (Fig. 4a), we constructed the fully connected graph G on a translated sample D of a continuous linear-structured metric space (X, d_X). Due to the absence of curvature, the metric space $(V(G), d_G)$ well-approximates (X, d_X) in the sense of an ϵ-correspondence (we omit an actual value of ϵ as we believe the concept is clear). Since G is fully connected, one connected component will be born in the filtration, and it will never die. This is illustrated by the persistence diagram in Fig. 4b, where we defined the filtration through the negative eccentricity function of G. Both for the ground truth model, as well as for G, the eccentricity function provides a smooth transition from the (underlying) leaves towards the center. However, the sublevel filtration for (X, d_X) will start at two connected components, that only merge at the center of X.

The second example (Fig. 4c) illustrates a 'finer' approximation of (X, d_X) through the Rips graph $R_{0.1}(D) := \mathrm{VR}^1_{0.1}(D)$ constructed from D. We now defined a function f (resp. g) on X (resp. D) that values 1 at every single point, apart from one point near the center where it values 1. Again, the filtration for $R_{0.1}(D)$ starts with one connected component (including all but one point), that never dies. The filtration for the ground truth model starts off with two connected components that merge only at the center as before.

The takeaway of the examples above, is that to ensure stability, we need two things. First, we need to formalize how well our graph G approximates the topology of the underlying space, both through the concept of ϵ-correspondences, as well as through a distance measure between nodes connected through an edge. Given a weighting function $w : E(G) \to \mathbb{R}^+$, we will use the maximum weight $w_{\max} := \max_{e\in E(G)} w(e)$ for this purpose. In practice, $w_{\max}$ will be low if the data is sufficiently densely sampled and G is a neighborhood graph. Second, the functions used to define the filtration must be such that if ϵ and $w_{\max}$ are small, so are a and b from Theorem 2.1. Inspired by Lemma 3.3 by Carrière et al. [9], we will consider *Lipschitz* functions, where a real-valued function f on a metric space (X, d) is called *c-Lipschitz* if $|f(x) - f(y)| \leq c d(x, y)$.

2.2. A new stability result for metric trees

In this section, we provide two closely-related functions to ensure stability for tree-structured topologies through graph approximations. These will be the (negative) *eccentricity* and the *normalized centrality*, the latter of which is scale-independent. The true persistence diagrams for these functions are extremely informative for metric trees. The birth of a component will always occur through a leaf, and its death through either a multifurcation or the center of the tree (Fig. 3d).

Definition 2.1. A *metric tree* is a path metric space (X, d_X) that is homeomorphic to a 1-dimensional stratified space, for which there is a unique path between every two points. The *radius* of X is $\mathrm{rad}(X) := \min_{x\in X} \max_{y\in X} d_X(x, y)$.

Theorem 2.2. *Let (X, d_X) be a metric tree, and G a positively weighted graph such that there exists an ϵ_X-correspondence C between (X, d_X) and (G, d_G). Let $f : X \to \mathbb{R}$, $g : V(G) \to \mathbb{R}$, and $\epsilon_f \in \mathbb{R}_{\geq 0}$ be such that for all $(x, u) \in C$, $|f(x) - g(u)| \leq \epsilon_f$, and f is c-Lipschitz. Then*

$$d_{\mathrm{b}}\Big(\mathrm{Dgm}_0\big(\overline{\mathcal{F}_f}(X)\big), \mathrm{Dgm}_0\big(\overline{\mathcal{F}_g}(G)\big)\Big) \leq c \max\left\{\frac{\epsilon_X}{2}, w_{\max}\right\} + c\epsilon_X + \epsilon_f.$$

Proof. Since the functional distortion ϵ_f and Lipschitz constant c remain the same after negating both functions, it suffices to show that the inequality holds for $-f$ and $-g$.

Take any $(x, u), (y, v) \in C$, let $P_{x,y} \subseteq X$ denote the unique path from x to y in X, and let $(u = p_0, p_1, \ldots, p_l = v)$ be a shortest path from u to v in G. For any $0 \leq i \leq l$, take q_i such that $(q_i, p_i) \in C$, with $q_0 = x, q_l = y$. Now arbitrarily take $t \in \mathbb{R}$.

Suppose first that $x \sim y$ in $\{z \in X : t \leq f(z)\}$. Let m_i be the closest point from q_i on $P_{x,y}$. If for any i, $d_X(q_i, m_i) > \frac{3\epsilon_X}{2}$, then

$$\begin{aligned} d_X(x, y) &= d_X(x, q_i) + d_X(q_i, y) - 2d_X(q_i, m_i) \\ &< d_X(x, q_i) + d_X(q_i, y) - 3\epsilon_X \leq d_G(u, v) - \epsilon_X \leq d_X(x, y), \end{aligned}$$

a contradiction. Now since necessarily $m_i \in \{z \in X : t \leq f(z)\}$,

$$g(p_i) \geq f(q_i) - \epsilon_f \geq f(m_i) - c d_X(m_i, q_i) - \epsilon_f \geq t - \frac{3c\epsilon_X}{2} - \epsilon_f.$$

This shows that $u \sim v$ in $G\Big[\Big\{w \in V(G) : t - \frac{3c\epsilon_X}{2} - \epsilon_f \leq g(w)\Big\}\Big]$.

Now suppose we have $x \nsim y$ in $\{z' \in X : t \leq f(z')\}$. If $x = y$, then $\max\{g(u), g(v)\} < t + \epsilon_f$, and $u \nsim v$ in $G\big[\big\{w \in V(G) : t + c(w_{\max} + \epsilon_X) + \epsilon_f \leq g(w)\big\}\big]$ (they are not included). If $x \neq y$, take any $z \in P_{x,y}$ that minimizes $f(z)$ over $P_{x,y}$. Observe that necessarily $f(z) < t$. Now let

$$i := \max\left\{0 \leq i < l : P_{q_i, P_{x,y}} \cap P_{z,y} = \emptyset \vee z = q_i\right\},$$

where $P_{q_i, P_{x,y}} \subseteq X$ denotes the unique path from q_i to (its closest point on) $P_{x,y}$ in X. It follows that

$$\begin{aligned} g(p_i) &\leq f(q_i) + \epsilon_f \leq f(z) + c d_X(q_i, z) + \epsilon_f \\ &\leq f(z) + c d_X(q_i, q_{i+1}) + \epsilon_f \\ &\leq f(z) + c(w_{\max} + \epsilon_X) + \epsilon_f < t + c(w_{\max} + \epsilon_X) + \epsilon_f. \end{aligned}$$

Again $u \nsim v$ in $G\big[\big\{w \in V(G) : t + c(w_{\max} + \epsilon_X) + \epsilon_f \leq g(w)\big\}\big]$. The result now follows from Theorem 2.1. □

Remark 2.1. The proof of Theorem 2.1 suggests that we can obtain even stronger comparisons by looking at the *interleaving distance* between the resulting merge trees, instead of the 0-dimensional persistence diagrams. Indeed, Morozov et al. [20] provide an example of two distinct merge trees for which the corresponding functions have the exact same persistence diagram. Unfortunately, computing interleaving distances between merge trees is currently computationally more challenging than computing bottleneck distances between persistence diagrams [26].

Remark 2.2. For Rips graphs $G = R_{3\delta}(D)$, the bound in Theorem 2.2 reduces to the bound in Lemma 3.3 by Carrière et al. [9] for zeroth-order persistent homology, whenever $\frac{\epsilon_X}{2} \leq w_{\max} \leq 3\delta$. However, our result applies to any graph, and does not require that $w_{\max}$ dominates $\frac{\epsilon_X}{2}$. Intuitive examples for which this is important include minimum spanning trees.

The convexity radius $\rho(X)$ states that for any open metric ball in X of radius less than $\rho(X)$, any two points x, y in this ball are connected by a unique shortest path on X. Similar to Lemma 3.3 by Carrière et al. [9], we expect that our result can be generalized to arbitrary length spaces by bounding ϵ_X through a function of the *convexity radius* $\rho(X)$ of X.

The following can now be straightforwardly derived.

Corollary 2.1. *Let (X, d_X) be a metric tree, and G a positively weighted graph such that there exists an ϵ-correspondence C between (X, d_X) and (G, d_G). Let $\mathcal{E}_X := \max_{x\in X} d_X(\cdot, x)$ be the eccentricity function, and $\mathscr{C}_X := \frac{\mathcal{E}_X^{\max} - \mathcal{E}_X(\cdot)}{\mathcal{E}_X^{\max} - \mathcal{E}_X^{\min}}$ be the normalized centrality function on X (define $\mathcal{E}_G$ and $\mathscr{C}_G$ analogously). Then*

$$d_{\mathrm{b}}\Big(\mathrm{Dgm}_0\big(\overline{\mathcal{F}_{-\mathcal{E}_X}}(X)\big), \mathrm{Dgm}_0\big(\overline{\mathcal{F}_{-\mathcal{E}_G}}(G)\big)\Big) \leq \max\left\{\frac{\epsilon}{2}, w_{\max}\right\} + 2\epsilon,$$

and

$$d_{\mathrm{b}}\Big(\mathrm{Dgm}_0\big(\overline{\mathcal{F}_{\mathscr{C}_X}}(X)\big), \mathrm{Dgm}_0\big(\overline{\mathcal{F}_{\mathscr{C}_G}}(G)\big)\Big) \leq \frac{\max\left\{\frac{\epsilon}{2}, w_{\max}\right\} + 5\epsilon}{\mathrm{rad}(X)},$$

where the last inequality holds if $\mathscr{C}_X$ and $\mathscr{C}_G$ are well-defined.

3. Experiments

In this section, we show how Theorem 2.2 can be applied in practice. We first illustrate this through synthetic data sampled from metric trees in Section 3.1. In Section 3.2, we provide novel insights and quality measures to the field of cell trajectory inference.

3.1. Synthetic data of metric trees

We considered four tree-structured topologies embedded in $\mathbb{R}^2$, and sampled 600 observations from each of them, by sampling uniformly from each branch a number of points proportional the length of this branch. For each of these data sets, we applied a

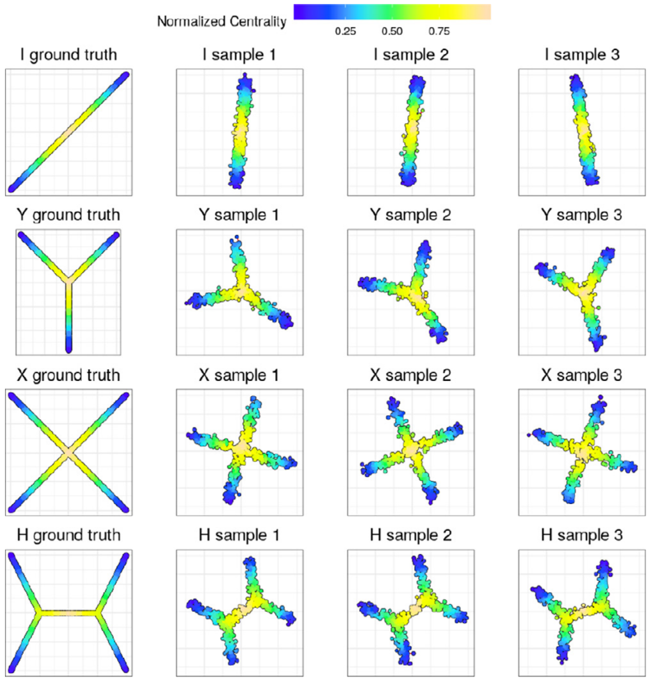


Fig. 5. Synthetic data sampled from the metric trees in the first column. The samples and their (MST) normalized centralities are shown in columns 2–4.

small amount of random 2-dimensional Gaussian noise, as well as a random rotation, three times. From each of these twelve resulting data sets, we constructed a Euclidean minimum spanning tree (MST), and computed the normalized centrality function. The resulting functions, MSTs, as well as the ground truth models, are shown in Fig. 5.

The persistence diagrams obtained for the sublevel filtrations of the normalized centrality functions are shown in Fig. 6. Note that there may be overlapping points. As can be expected, there are many points in the persistence diagrams for the MSTs near the diagonal. This is a result from the MST not including any triangles (in the graph theoretical sense). Nevertheless, we observe that the highly elevated points in all our diagrams identify important structural information of the ground truth models.

Fig. 7 a visualizes the pairwise bottleneck distances between all diagrams. Fig. 7b shows a Multi-Dimensional Scaling (MDS) plot of this distance matrix. We see that similar shapes are clustered well together. We also note that the H-structured topologies are somewhat in the middle of the other topologies. This is as expected. E.g., the longer the middle branch of the corresponding model is, the closer this pattern is to a I-pattern. The shorter this branch is, the closer it is to an X-pattern.

3.2. Cell trajectory data

Cell trajectory inference considers the task of inferring a graph-structured model from *gene expression data*, to identify the differentiation process of the cells. Cells can be regarded as points in a (high-dimensional gene expression) space $\mathbb{R}^d$, and approximate (the embedding of) their underlying graph-structured model in this space. Some examples of cell trajectory data sets and their underlying models are illustrated in Fig. 9.

Cell trajectory inference is overall a very difficult task. Even the top ranked methods have a low performance on many data sets [25]. The purpose of this section is not to propose the use of our signatures (Corollary 2.1) as a new topological inference method for this type of data, but rather to use these to study why this problem is essentially so difficult. In particular, Vandaele et al. [28] recently showed that state-of-the-art cell trajectory inference methods struggle to approximate the geometry of the underlying model well, or commonly underestimate the number of leaves. To explain these difficulties, we proceed with an analysis similar to the one in Section 3.1.

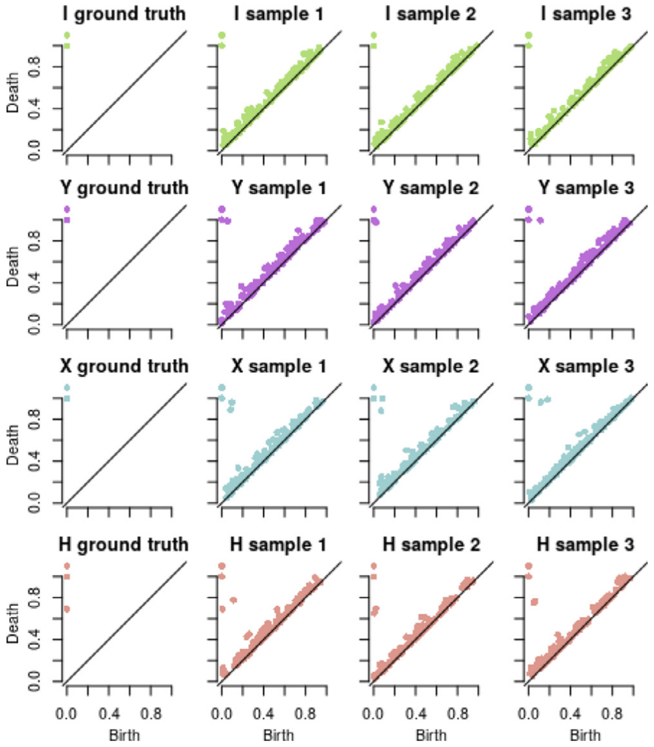
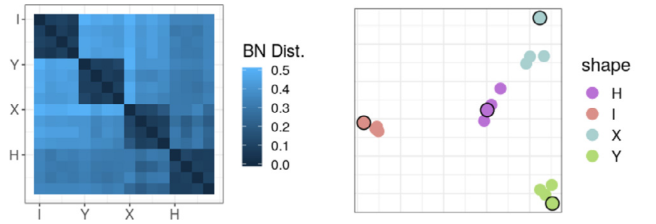


Fig. 6. The ground truth and empirical persistence diagrams are computed using the normalized centrality to define the filtration.



(a) Pairwise bottleneck distances between all our true and experimental diagrams. The ground truths are marked by their corresponding shape.

(b) MDS plot of the pairwise bottleneck distances. The points corresponding to the ground truth models are marked by a black contour.

Fig. 7. Visualizing the bottleneck distances between the diagrams.

We consider 131 synthetic and 57 real cell trajectory data sets with an underlying tree-structured model [5]. The number of cells ranged from 59 to 5018, and the number of genes from 373 to 23,658. A two-dimensional diffusion map embedding was computed for each data set, both for visualization purposes, as well as to reduce the effects of the *curse of dimensionality* on our neighborhood graph approximation [24]. A 10NN graph and its normalized centralities were computed from each embedding.

Fig. 8 visualizes all cell trajectory data sets by means of an MDS plot of the pairwise bottleneck distances we obtained through topological persistence of our 10NN graphs. We illustrate twelve 'landmark' embeddings of cell trajectory data sets, as well as their ground truth models on these embeddings, and their obtained empirical persistence diagrams in Fig. 9.

First, observe that all linear cell trajectories are located near a linear curve on top of the MDS plot. This means that our chosen

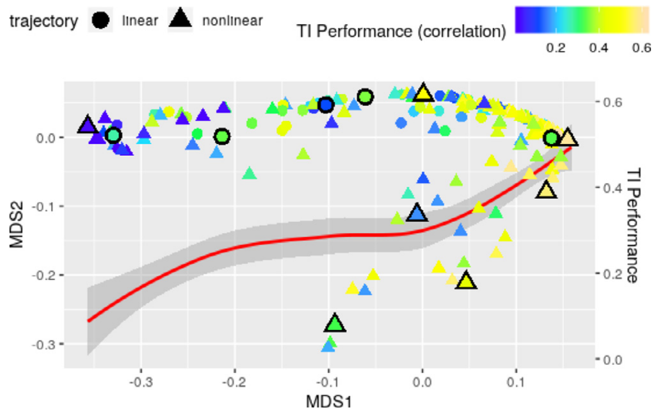


Fig. 8. MDS plot of pairwise bottleneck distances of the persistence diagrams obtained through the 10NN graphs and normalized centralities. Each point corresponds to one cell trajectory data set. A loess curve (red) is fitted using the MDS1 coordinate as independent variable, and the average performance over all considered cell trajectory inference methods as dependent variable. The points with a black contour correspond to the data sets visualized in Fig. 9. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

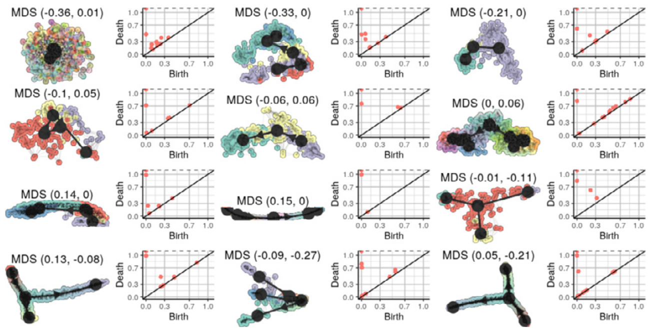


Fig. 9. Twelve example data sets and their corresponding empirical persistence diagram. The coloring corresponds to the ground truth grouping of cells.

data representation does not artificially create more leaves than truthfully present. E.g., this is more often the case when we apply a PCA projection instead of a diffusion map embedding. However, many nonlinear trajectories are located near this curve as well. Near the right side of this curve, this is mainly due to branches being relatively short compared to a main linear trajectory (e.g., MDS (0, 0.6) in Fig. 9). These trajectories are indeed theoretically close to linear according to our chosen metric. On the left side of this curve, we find the more noisy data sets, where we fail to provide a good representation. Their persistence diagrams represent more ‘blob’-like patterns (Fig. 9). Below this curve, we find the trajectories where we truthfully manage to identify additional branches. However, we note that it appears to be difficult to identify more than three leaves. This explains why cell trajectory inference methods commonly underestimate the true number of leaves [28]. Note that the ‘boomerang’ shape made up by all data sets in Fig. 8, also coincides with what we theoretically expect for our chosen metric. We note a continuous transmission of blob-like patterns, towards linear patterns, towards patterns with leaves. The fact that this shape takes a turn near the right, can be theoretically explained through the definition of the bottleneck distance. As we only look at the maximal distances of a matching, the number of ‘high’ distances in such matching does not matter. Blob-like patterns are as distant from linear patterns as they are from patterns with more leaves, according to this metric.

Finally, we fitted a loess curve (standard settings in R) using the MDS1 coordinate as the independent and the average performance over 45 different cell trajectory inference methods as the dependent variable. This performance is measured through the geodesic distance preservation (correlation) metric introduced by Saelens et al. [25]. Fig. 8 shows a positive correlation (0.58) between these variables. Note that the choice of using the MDS1 coordinate is arbitrary in general. However, this choice supports our findings that on the left side of our MDS plot, we mainly find noisy data sets. Since every cell trajectory inference method uses a different algorithm or data representation (such as the type of dimensionality reduction or neighborhood graph), this can be seen as a quality measure of the data itself, independent of our chosen data representation.

4. Discussion and conclusion

We provided a novel foundation for quantifying topological patterns in metric trees through graph approximations, which led to new and direct stability results. Though these result currently only holds for metric trees, we opened up new possibilities to study which functions ensure stability by means of Theorems 2.1, 2.2, and Remark 2.2. This may lead to further theoretical justification of recognizing a wider variety of patterns through graph approximations.

Rather than using our signatures for topological inference, we introduced a novel use for them in an exploratory data analysis setting. We developed insights into cell trajectory inference, and provided the first charting of such data sets that explains some of the difficulties this field is confronted with. We also provided a new way of quality measurement, that does not require ground truth knowledge. It will be interesting to investigate whether other types of signatures, such as those discussed in Section 1.3, may find additional applications within this setting.

Since we consider sublevel filtrations on any given graph, we can choose to approximate our data through a small or sparse graph. For 0-dimensional persistence, this is computationally more efficient than Rips based signatures for metric spaces equipped with functions, which may require the construction of the complete graph on the data to include all nodes. Nevertheless, for larger data sets it may be interesting to explore approximation methods similar to the *witness complexes* for Rips based filtrations [12].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the ERC under the EU’s Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no. 615517 and under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 963924), by the Flemish government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme, and by the FWO (project no. G091017N, G0F9816N, 3G042220).

References

[1] M. Aanjaneya, F. Chazal, D. Chen, M. GLisse, L. Guibas, D. Morozov, Metric graph reconstruction from noisy data, Int. J. Comput. Geom. Appl. 22 (04) (2012) 305–325.
[2] R. Assaf, A. Goupil, V. Vrabie, T. Boudier, M. Kacim, Persistent homology for object segmentation in multidimensional grayscale images, Pattern Recognit. Lett. 112 (2018) 277–284.
[3] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, D. Morozov, Inferring local homology from sampled stratified spaces, in: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), IEEE, 2007, pp. 536–546.

[4] S.H.-D. Boissonnat, An efficient implementation of Delaunay triangulations in medium dimensions (2008).

[5] R. Cannoodt, W. Saelens, H. Todorov, Y. Saeys, Single-cell -omics datasets containing a trajectory, 2018, 10.5281/zenodo.1443566.

[6] G. Carlsson, Topology and data, Bull. Am. Math. Soc. 46 (2) (2009) 255–308.

[7] G. Carlsson, Topological pattern recognition for point cloud data, Acta Numer. 23 (2014) 289368.

[8] M. Carriere, B. Michel, S. Oudot, Statistical analysis and parameter selection for mapper, J. Mach. Learn. Res. 19 (1) (2018) 478–516.

[9] M. Carrière, S. Oudot, M. Ovsjanikov, Local signatures using persistence diagrams(2015).

[10] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Mémoli, S.Y. Oudot, Gromov–Hausdorff stable signatures for shapes using persistence, in: Computer Graphics Forum, 28, Wiley Online Library, 2009, pp. 1393–1403.

[11] F. Chazal, V. de Silva, S. Oudot, Persistence stability for geometric complexes, Geom. Dedicata 173 (2014) 193–214.

[12] V. De Silva, G.E. Carlsson, Topological estimation using witness complexes, SPBG 4 (2004) 157–166.

[13] T.K. Dey, F. Mémoli, Y. Wang, Mutiscale mapper: a framework for topological summarization of data and maps, arXiv preprint arXiv:1504.03763(2015).

[14] B. Di Fabio, C. Landi, Persistent homology and partial similarity of shapes, Pattern Recognit. Lett. 33 (11) (2012) 1445–1450.

[15] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, Applied Mathematics, American Mathematical Society, 2010.

[16] R. Ghrist, Barcodes: the persistent topology of data, Bull. (New Series) Am. Math. Soc. 45 (107) (2008) 61–75.

[17] C. Giusti, R. Ghrist, D.S. Bassett, Two's company, three (or more) is a simplex, J. Comput. Neurosci. 41 (1) (2016) 1–14.

[18] A. Hatcher, Algebraic Topology, Cambridge University Press, 2002.

[19] X. Liu, Z. Xie, D. Yi, et al., A fast algorithm for constructing topological structure in large data, Homol. Homotopy Appl. 14 (1) (2012) 221–238.

[20] D. Morozov, K. Beketayev, G. Weber, Interleaving distance between merge trees, Discrete Comput. Geom. 49 (2013) 22–45.

[21] M. Nicolau, A.J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, Proc. Natl. Acad. Sci. 108 (17) (2011) 7265–7270.

[22] S.Y. Oudot, Persistence Theory: From Quiver Representations to Data Analysis, Mathematical Surveys and Monographs, American Mathematical Society, 2015.

[23] V. Patrangenaru, L. Ellingson, Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis, 1first ed., CRC Press, Inc., USA, 2015.

[24] M. Radovanović, A. Nanopoulos, M. Ivanović, Nearest neighbors in high-dimensional data: the emergence and influence of hubs, in: Proceedings of the 26th Annual International Conference on Machine Learning, in: ICML '09, ACM, New York, NY, USA, 2009, pp. 865–872.

[25] W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods, Nat. Biotechnol. 37 (2019) 1.

[26] E.F. Touli, Y. Wang, FPT-algorithms for computing Gromov–Hausdorff and interleaving distances between trees, arXiv preprint arXiv:1811.02425(2018).

[27] R. Vandaele, G.A. Nervo, O. Gevaert, Topological image modification for object detection and topological image processing of skin lesions, Sci. Rep. 10 (1) (2020) 1–15.

[28] R. Vandaele, Y. Saeys, T.D. Bie, Mining topological structure in graphs through forest representations, J. Mach. Learn. Res. 21 (215) (2020) 1–68.