THE RELATIONSHIP BETWEEN THE INTRINSIC ČECH AND PERSISTENCE DISTORTION DISTANCES FOR METRIC GRAPHS

Ellen Gasparovic,[*] *Maria Gommel,*[†] *Emilie Purvine,*[‡] *Radmila Sazdanovic,*[§] *Bei Wang,*[¶] *Yusu Wang,*[‖] *and Lori Ziegelmeier*[**]

Abstract. Metric graphs are meaningful objects for modeling complex structures that arise in many real-world applications, such as road networks, river systems, earthquake faults, blood vessels, and filamentary structures in galaxies. To study metric graphs in the context of comparison, we are interested in determining the relative discriminative capabilities of two topology-based distances between a pair of arbitrary finite metric graphs: the persistence distortion distance and the intrinsic Čech distance. We explicitly show how to compute the intrinsic Čech distance between two metric graphs based solely on knowledge of the shortest systems of loops for the graphs. Our main theorem establishes an inequality between the intrinsic Čech and persistence distortion distances in the case when one of the graphs is a bouquet graph and the other is arbitrary. The relationship also holds when both graphs are constructed via wedge sums of cycles and edges.

1 Introduction

When working with graph-like data equipped with a notion of distance, a very useful means of capturing existing geometric and topological relationships within the data is via a *metric graph*. Given an ordinary graph $G = (V, E)$ and a length function on the edges, one may view G as a metric space with the shortest path metric in any geometric realization.

Metric graphs are used to model a variety of real-world data sets, such as road networks, river systems, earthquake faults, blood vessels, and filamentary structures in galaxies [1, 25, 26]. Given these practical applications, it is natural to ask how to compare two metric graphs in a meaningful way. Such a comparison is important to understand the stability of these structures in a noisy setting. One way to do this is to check whether there is a bijection between the two input graphs as part of a graph isomorphism problem, which is NP-complete [3]. Another way is to define, compute, and compare various distances on

[*] *Department of Mathematics, Union College*, gasparoe@union.edu

[†] *Department of Mathematics, University of Iowa*, maria-gommel@uiowa.edu

[‡] *Computing and Analytics Division, Pacific Northwest National Laboratory*, Emilie.Purvine@pnnl.gov

[§] *Department of Mathematics, North Carolina State University*, rsazdanovic@math.ncsu.edu

[¶] *School of Computing and Scientific Computing and Imaging Institute, University of Utah*, beiwang@sci.utah.edu

[‖] *Department of Computer Science, Ohio State University*, yusu@cse.ohio-state.edu

[**] *Department of Mathematics, Statistics, and Computer Science, Macalester College*, lziegel1@macalester.edu

the space of graphs. In this paper, we are interested in determining the discriminative capabilities of two distances (both of which may be defined on arbitrary metric spaces, not just metric graphs) that arise from computational topology: the persistence distortion distance and the intrinsic Čech distance. If two distances d_1 and d_2 on the space of metric graphs satisfy an inequality $d_1(G_1, G_2) \leq c \cdot d_2(G_1, G_2)$ (for some constant $c > 0$ and any pair of metric graphs G_1 and G_2), this means that d_2 has greater discriminative capacity for differentiating between two input graphs. For instance, if $d_1(G_1, G_2) = 0$ and $d_2(G_1, G_2) > 0$, then d_2 has a better discriminative power than d_1.

1.1 Related work

Well-known methods for comparing graphs using distance measures include combinatorial (e.g., graph edit distance [28]) and spectral (e.g., eigenvalue decomposition [27]) approaches. Graph edit distance minimizes the cost of transforming one graph to another via a set of elementary operators such as node/edge insertions/deletions, while spectral approaches optimize objective functions based on properties of the graph spectra.

Recently, several distances for comparing metric graphs have been proposed based on ideas from computational topology. In the case of a special type of metric graph called a Reeb graph, these distances include: the functional distortion distance [4], the combinatorial edit distance [15], the interleaving distance [23], and its variant in the setting of merge trees [19]. In particular, the functional distortion distance can be considered as a variation of the Gromov-Hausdorff distance between two metric spaces [4]. The interleaving distance is defined via algebraic topology and utilizes the equivalence between Reeb graphs and cosheaves [23]. For metric graphs in general, both the persistence distortion distance [13] and the intrinsic Čech distance [11] take into consideration the structure of metric graphs, independent of their geometric embeddings, by treating them as continuous metric spaces. In [21], Oudot and Solomon point out that since compact geodesic spaces can be approximated by finite metric graphs in the Gromov–Hausdorff sense [6] (see also the recent work of Mémoli and Okutan [18]), one can study potentially complicated length spaces by studying the persistence distortion of a sequence of approximating graphs.

In the context of comparing the relative discriminative capabilities of these distances, Bauer, Ge, and Wang [4] show that the functional distortion distance between two Reeb graphs is bounded from below by the bottleneck distance between the persistence diagrams of the Reeb graphs. Bauer, Munch, and Wang [5] establish a strong equivalence between the functional distortion distance and the interleaving distance on the space of all Reeb graphs, which implies the two distances are within a constant factor of one another. Carrière and Oudot [9] consider the intrinsic versions of the aforementioned distances and prove that they are all globally equivalent. They also establish a lower bound for the bottleneck distance in terms of a constant multiple of the functional distortion distance. In [13], Dey, Shi, and Wang show that the persistence distortion distance is stable with respect to changes to input metric graphs as measured by the Gromov-Hausdorff distance. In other words, the persistence distortion distance is bounded above by a constant factor of the Gromov-Hausdorff distance. Furthermore, the intrinsic Čech distance is also bounded from above by the Gromov-Hausdorff distance for general metric spaces [11].

1.2 Our contribution

The main focus of this paper is relating two specific topological distances between general metric graphs G_1 and G_2: the intrinsic Čech distance and the persistence distortion distance. Both of these can be viewed as distances between topological signatures of G_1 and G_2. Indeed, in the case of the intrinsic Čech distance, a metric graph (G, d_G) is mapped to the persistence diagram $\mathrm{Dg}_1 IC_G$ induced by the so-called intrinsic Čech filtration IC_G, and we may think of $\mathrm{Dg}_1 IC_G$ as the signature of G. The intrinsic Čech distance $d_{IC}(G_1, G_2)$ between two metric graphs G_1 and G_2 is the bottleneck distance between these signatures, denoted $d_B(\mathrm{Dg}_1 IC_{G_1}, \mathrm{Dg}_1 IC_{G_2})$.

For the persistence distortion distance, each metric graph G is mapped to a set $\Phi(G)$ of persistence diagrams, which is the signature of the graph G in this case. The persistence distortion distance $d_{PD}(G_1, G_2)$ between G_1 and G_2 is measured by the Hausdorff distance between these signatures. See Section 2 for the definition of Φ, along with more detailed definitions of these two distances.

Our objective is to determine the relative discriminative capacities of such signatures. We conjecture that the persistence distortion distance is more discriminative than the intrinsic Čech distance.

Conjecture 1. *$d_{IC} \leq c \cdot d_{PD}$ for some constant $c > 0$.*

It is known from [16] that $\mathrm{Dg}_1 IC_G$ depends only on the lengths of the shortest system of loops in G, and thus, the persistence distortion distance appears to be more discriminative, intuitively. We show in Section 3 that the intrinsic Čech distance between two arbitrary finite metric graphs is determined solely by the difference in these shortest cycle lengths; see Theorem 5 for a precise statement. This further implies that the intrinsic Čech distance between two arbitrary metric trees is always 0. In contrast, the persistence distortion distance takes relative positions of loops as well as branches into account, and is nonzero in the case of two trees. In other words, the conjecture holds for metric trees.

We make progress toward proving the conjecture in greater generality in this paper. Theorem 11 establishes an inequality between the intrinsic Čech and persistence distortion distances for two finite metric graphs in the case when one of the graphs is a bouquet graph and the other is arbitrary. In this case, the constant is $c = 1/2$. The theorem and proof appear in Section 4, and we conclude that section by proving that Conjecture 1 also holds when both graphs are constructed by taking wedge sums of cycles and edges. We believe there are potential applications to the study of RNA foldings in the case when both graphs are obtained via wedges of cycles and edges.

While this does not yet prove the conjecture for arbitrary metric graphs, our work provides the first non-trivial relationship between these two meaningful topological distances. Our proofs also provide insights on the map Φ from a metric graph into the space of persistence diagrams as utilized in the definition of the persistence distortion distance. This map Φ is of interest itself; indeed, see the recent study of this map in [21].

In general, we believe that this direction of establishing qualitative understanding of topological signatures and their corresponding distances is interesting and valuable for use in applications. We leave the proof of the conjecture for arbitrary metric graphs as an open problem, discuss its technical challenges, and give a brief discussion on some future directions in Section 5.

2 Background

2.1 Persistent homology and metric graphs

We begin with a brief summary of *persistent homology* and how it can be utilized in the context of *metric graphs*. We specifically focus on *extended persistence*, defined in [12, 14], as it is crucial to the definition of the involved distances. For background on homology and simplicial complexes, we refer the reader to [17, 20], and for further details on persistent homology, see, e.g., [7, 14].

Persistent homology. Consider a continuous function $f : X \to \mathbb{R}$ on a compact topological space X and suppose f is a Morse function, or more generally, f is of *Morse type* [10]. Let $a_1 < a_2 < \ldots < a_n$ be the critical values of f and $b_0 < a_1 < b_1 < a_2 < \ldots < a_n < b_n$ be the interleaved regular values. Define *sublevel sets* $X_i = f^{-1}(-\infty, b_i]$ and *superlevel sets* $X^i = f^{-1}[b_i, \infty)$. By compactness, we have $X_0 = X^n = \emptyset$, $X_n = X^0 = X$. Extended persistence arises from the following sequence of absolute and relative homology groups connected by inclusions (referred to as an *extended filtration*) in any homological dimension k with coefficients in some field:

$$\begin{aligned} 0 = H_k(X_0) &\to \ldots \to H_k(X_n) \\ &\to H_k(X, X^n) \to \ldots \to H_k(X, X^0) = 0. \end{aligned}$$

Elements of each homology group are tracked through the above filtration. In the top row of the filtration, the inclusions $X_i \hookrightarrow X_j$ induce maps $f^{i,j} : H(X_i) \to H(X_j)$ for $0 \leq i < j \leq n$. A homological element (class) $\alpha \in H(X_i)$ is *born* at X_i if $\alpha \notin \mathrm{im} f^{i-1,i}$; it *dies* entering X_j if $f^{i,j}(\alpha) \in \mathrm{im} f^{i-1,j}$ but $f^{i,j-1}(\alpha) \notin \mathrm{im} f^{i-1,j-1}$. The notion of birth and death can be extended to the bottom row as well as the entire filtration, giving rise to an *extended persistence diagram*. An (extended) persistence diagram is a multi-set of points in the plane $(\mathbb{R} \cup \pm\infty)^2$: it contains a point (a_i, a_j) for each element that is born at X_i (or at (X, X^i)) and dies at X_j (or at (X, X^j)). An *ordinary persistence point* arises from an element that is born and dies in the top row, a *relative persistence point* arises from an element that is born and dies in the bottom row, and an *extended persistence point* corresponds to an element that is born in the top row and dies in the bottom row. It has been shown in [8], via the Mayer–Vietoris pyramid, that the 0- and 1-dimensional extended persistence diagrams of f are equivalent to the 0-dimensional *levelset zigzag persistence diagram* induced by f. In this paper, we restrict our attention to only the 1-dimensional extended persistence points, as these correspond to loops inherent in the underlying structure of a metric graph.

We illustrate the idea of extended persistence with an example in Figure 1. We equip a topological space on the left with a height function, whose critical values are specified

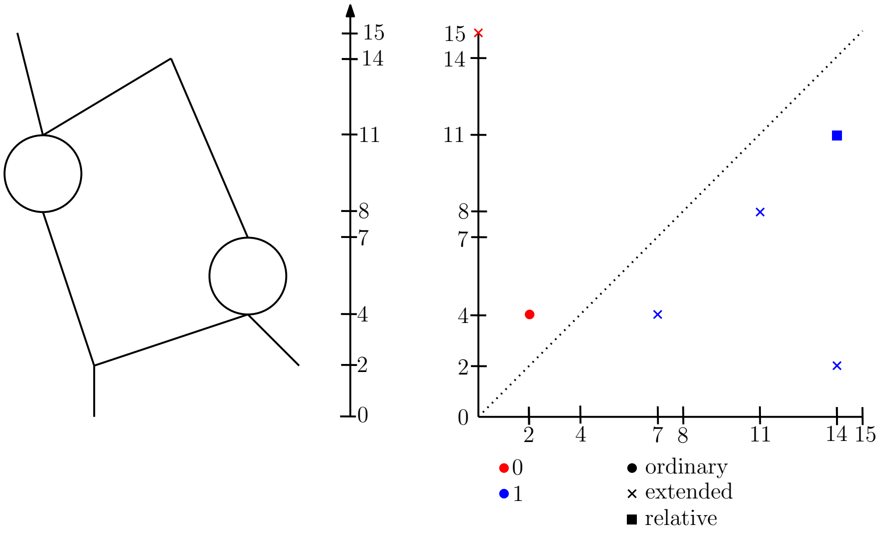

Figure 1: A topological space and its associated height function (left) and its corresponding extended persistence diagram (right). Red indicates 0-dimensional and blue indicates 1-dimensional features. Circles, crosses, and squares are ordinary, extended and relative persistence points, respectively.

vertically. Its corresponding extended persistence diagram is shown on the right. We track the evolution of homological elements following the above extended filtration. In the top row of the filtration, we see that two 0-dimensional elements (connected components) are born at 0 and 2, respectively. The component that is born at 2 dies at 4 by merging with the component that is born at 0. This event gives rise to an ordinary persistence point $(2, 4)$. We also see that two smaller loops are born at 7 (right loop) and 11 (left loop) respectively, while a large loop is not born until 14. In the bottom row of the filtration, from left to right, we observe that the 0-dimensional element that is born at 0 now dies at 14 via the relative homology computation. Each of the loops born in the first row now dies in the second row: the large loop dies at 2, the small loop on the right dies at 4, and the small loop on the left dies at 8. These events correspond to four extended persistence points, $(0, 14)$ (in dimension 0), $(7, 4)$, $(11, 8)$, and $(14, 2)$ (in dimension 1). Finally, one additional loop is born at 14 via the relative homology, which dies at 11 and gives rise to a relative persistence point $(14, 11)$.

Metric graphs. We are interested in summarizing the topological information of a finite *metric graph*, specifically in homological dimension $k = 1$. Given a graph $G = (V, E)$, where V and E denote the vertex and edge sets, respectively, as well as a length function, $length : E \to \mathbb{R}_{\geq 0}$, on edges in E, a finite metric graph $(|G|, d_G)$ is a metric space where $|G|$ is a geometric realization of G and d_G is defined as in [13]. Namely, if e and $|e|$ denote an edge and its image in the geometric realization, we define $\alpha : [0, length(e)] \to |e|$ to be the arclength parametrization, so that $d_G(u,v) = |\alpha^{-1}(v) - \alpha^{-1}(u)|$ for any $u, v \in |e|$. This definition may then be extended to any two points in $|G|$ by restricting a given path from one point to another to edges in G, adding up these lengths, then taking the distance to be the minimum length of any such path. In this way, all points along an edge are points in a metric graph, not just the original graph's vertices.

A *system of loops of G* refers to a set of cycles whose associated homology classes form a minimal generating set for the 1-dimensional (singular) homology group of G. The *length-sequence* of a system of loops is the sequence of lengths of elements in this set listed in non-decreasing order. Thus, a system of loops of G is *shortest* if its length-sequence is lexicographically smallest among all possible systems of loops of G. Intuitively, these loops are the extended persistence features (not the ordinary or relative features) in 1-dimensional extended persistence, and as such, we restrict our discussion to these extended persistence points.

One particular class of metric graphs we will be working with are *bouquet graphs*. These are metric graphs containing a single vertex with a number of self-loops of various lengths attached to it.

2.2 Intrinsic Čech and persistence distortion distances

In this section, we recall the distances between metric graphs that are being explored in this work. We note that both are actually *pseudo-distances* because it can be the case that $d(G_1, G_2) = 0$ when $G_1 \neq G_2$. However, for ease of exposition, we will refer to them simply as distances in this paper. Both rely on the *bottleneck distance* on the space of persistence diagrams, a version of which we now state.

Definition 2. *Let X and Y be persistence diagrams with $\mu : X \to Y$ a bijection. The* ***bottleneck distance*** *between X and Y is*

$$d_B(X,Y) := \inf_{\mu: X \to Y} \sup_{x \in X} ||x - \mu(x)||_1.$$

Although this definition differs from the standard version of the bottleneck distance, which uses $||x - \mu(x)||_\infty$ rather than $||x - \mu(x)||_1$, the two are related via the inequalities $||x||_\infty \leq ||x||_1 \leq 2||x||_\infty$.

Next, let (G, d_G) be a metric graph with geometric realization $|G|$. Define the intrinsic ball $B(x, a_i) = \{y \in |G| : d_G(x,y) \leq a_i\}$ for any $x \in |G|$, as well as the uncountable open cover $U_{a_i} = \{B(x, a_i) : x \in |G|\}$. We use $\check{\text{C}}\text{ech}(a_i)$ to denote the nerve of the cover U_{a_i}, referred to as the *intrinsic Čech complex*. See Figure 2 for an illustration. Then $\{\check{\text{C}}\text{ech}(a_i) \hookrightarrow \check{\text{C}}\text{ech}(a_j)\}_{0 \leq a_i < a_j}$ is the *intrinsic Čech filtration* inducing the *intrinsic Čech persistence module* $\{H_k(\check{\text{C}}\text{ech}(a_i)) \to H_k(\check{\text{C}}\text{ech}(a_j))\}_{0 \leq a_i < a_j}$ in any dimension k, and the corresponding persistence diagram is denoted $Dg_k IC_G$. In this paper, we work with dimension $k = 1$. The intrinsic Čech complex was first introduced in [11]. The stability theorem that appears in that paper could be rephrased in terms of the *intrinsic Čech distance* defined below.

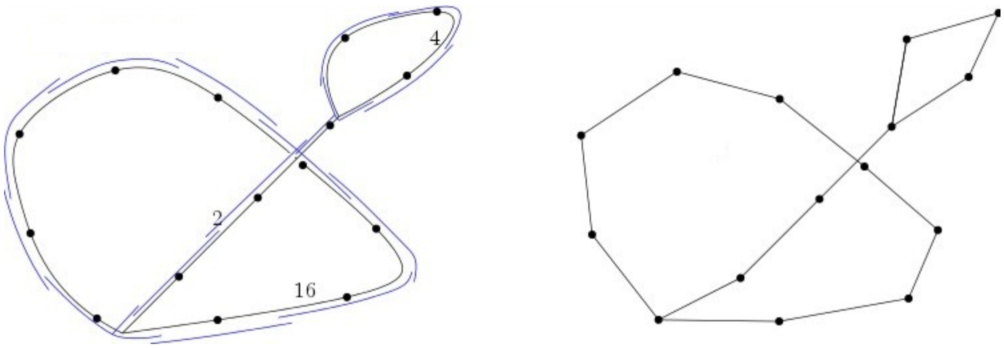

Figure 2: A finite subset of the infinite cover at a fixed radius (left) and its corresponding nerve (right).

Definition 3. *Given two metric graphs* (G_1, d_{G_1}) *and* (G_2, d_{G_2})*, their* ***intrinsic Čech distance*** *is*

$$d_{IC}(G_1, G_2) := d_B(\mathrm{Dg}_1 IC_{G_1}, \mathrm{Dg}_1 IC_{G_2}).$$

The persistence distortion distance was first introduced in [13]. Given a base point $v \in |G|$, define the geodesic distance function $f_v : |G| \to \mathbb{R}$ where $f_v(x) = d_G(v, x)$. In [13], $\mathrm{Dg}(f_v)$ is the union of the $0-$ and $1-$dimensional extended persistence diagrams for f_v, or equivalently, it is the $0-$dimensional levelset zigzag persistence diagram as noted in Section 2.1. In this paper, however, we take $\mathrm{Dg}(f_v)$ to be only the 1-dimensional extended persistence points as they correspond to loops in a metric graph for computing the persistence distortion signature. We therefore abuse notation by using $\mathrm{Dg}(f_v)$ to refer to the diagram containing only the 1-dimensional extended persistence points. Define $\Phi : |G| \to SpDg$, $\Phi(v) = \mathrm{Dg}(f_v)$, where $SpDg$ denotes the space of persistence diagrams for all points $v \in |G|$. The set $\Phi(|G|) \subset SpDg$ is the *persistence distortion* of the metric graph G.

Definition 4. *Given two metric graphs* (G_1, d_{G_1}) *and* (G_2, d_{G_2})*, their* ***persistence distortion distance*** *is*

$$d_{PD}(G_1, G_2) := d_H(\Phi(|G_1|), \Phi(|G_2|))$$

where d_H *denotes the Hausdorff distance. In other words,*

$$d_{PD}(G_1, G_2) = \max\left\{ \sup_{D_1 \in \Phi(|G_1|)} \inf_{D_2 \in \Phi(|G_2|)} d_B(D_1, D_2), \sup_{D_2 \in \Phi(|G_2|)} \inf_{D_1 \in \Phi(|G_1|)} d_B(D_1, D_2) \right\}.$$

We give an example in Figure 3 to illustrate the intuitive connection between the loops in a metric graph and the 1-dimensional extended persistence points. A graph G is given with v being the base point of a geodesic distance function $f_v : |G| \to \mathbb{R}$. For simplicity, assume all edges are of unit length. In this example, following the extended filtration using the sublevel and superlevel sets of f_v, we see that four loops are born in the top row of the filtration, at critical values 2, 3, 4.5, and 5.5, respectively. The birth time of each loop corresponds to a location (not necessarily a vertex) in the graph that is furthest away from the base point (i.e., a local maximum of f_v). Now, all four such loops die in the bottom row of the filtration. The death time of each loop corresponds to a location in the same loop whose distance is the closest to the base point. Thus, the loops born at 3, 4.5, and 5.5 all die at 1, while the loop born at 2 dies at 0. We therefore obtain four 1-dimensional extended persistence points: $(2, 0), (3, 1), (4.5, 1)$, and $(5.5, 1)$. The right of Figure 3 illustrates the general correspondence between loops in a metric graph and the critical points in $|G|$ that correspond to 1-dimensional extended persistence points.

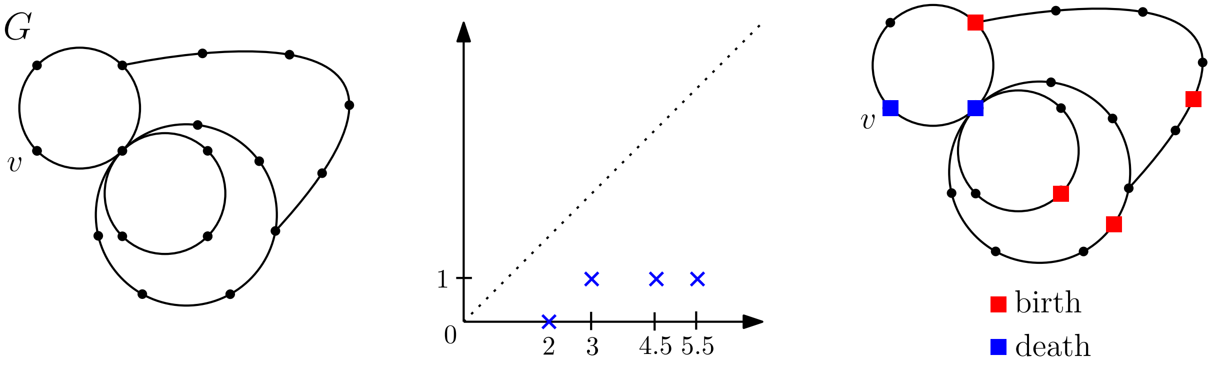

Figure 3: A metric graph G with a geodesic distance function $f_v : |G| \to \mathbb{R}$ from base point v where graph edges are unit length (left). This graph gives rise to four 1-dimensional extended persistence points: $(2, 0), (3, 1), (4.5, 1)$, and $(5.5, 1)$ (center). Red squares indicate locations (critical points of f_v) at which a cycle is born, while blue squares show where a cycle dies (right).

3 Calculating the intrinsic Čech distance

In this section, we show that the intrinsic Čech distance between two metric graphs may be easily computed from knowing the shortest systems of loops for the graphs. We begin with a theorem that characterizes the bottleneck distance between two sets of points in the extended plane.

Theorem 5. *Let* $D_1 = \{(0, a_1), \ldots, (0, a_n)\}$ *and* $D_2 = \{(0, b_1), \ldots, (0, b_n)\}$ *be two persistence diagrams with* $0 \leq a_1 \leq \cdots \leq a_n$ *and* $0 \leq b_1 \leq \cdots \leq b_n$*, respectively. Then* $d_B(D_1, D_2) = \max_{i=1}^{n} |a_i - b_i|$.

Proof. To simplify notation, we use the convention that for all $i = 1, \ldots, n$, $(0, a_i) = \overline{a_i}$, $(0, b_i) = \overline{b_i}$, and $(0, 0) = \overline{0}$. Let μ be any matching of points in D_1 and D_2, where each point $\overline{a_i}$ in D_1 is either matched to a unique point $\overline{b_j}$ in D_2 or to the nearest neighbor in the diagonal (and similarly for D_2). Assume that C_μ is the cost of the matching μ, i.e., the maximum distance between two matched points.

Now, let μ^* be the matching such that $\mu^*(\overline{a_i}) = \overline{b_i}$ for all $0 \leq i \leq n$. By construction, the cost of this matching is $C_{\mu^*} = \max_{i=1}^{n} |a_i - b_i|$. We claim that the matching cost of μ^* is less than or equal to that of μ, i.e., $C_{\mu^*} \leq C_\mu$. If this is the case, then μ^* is the optimal bottleneck matching and therefore $d_B(D_1, D_2) = C_{\mu^*}$.

To show this, we look at where the matchings μ and μ^* differ. Note that since all of the off-diagonal points in D_1 and D_2 lie on the y-axis, any such point matched to the diagonal under μ may simply be matched to $(0, 0)$ since this will yield the same value in the ℓ_1-norm. Now, starting with b_1, let j be the first index where $\mu(\overline{a_j}) \neq \overline{b_j}$. Then, we have two cases: (1) $\mu(\overline{a_k}) = \overline{b_j}$ for some $k > j$ (i.e., $\overline{b_j}$ is matched with some $\overline{a_k} \neq \overline{a_j}$); or (2) $\mu(\overline{0}) = \overline{b_j}$ (i.e., $\overline{b_j}$ is matched with the diagonal, or equivalently, to $\overline{0}$). We show that in either case, matching $\overline{b_j}$ with $\overline{a_j}$ instead does not increase the cost of the matching.

In the first case, let us also assume that $\mu(\overline{a_j}) = \overline{b_l}$ for some $l > j$ (the situation where $\mu(\overline{a_j}) = \overline{0}$ will be taken care of in the second case). Then, $\max\{|a_j - b_j|, |a_k - b_l|\} \leq \max\{|a_j - b_l|, |a_k - b_j|\}$. That is, if we were to instead pair $\overline{a_j}$ with $\overline{b_j}$ and $\overline{a_k}$ with $\overline{b_l}$, the cost of the matching would be lower. This can be seen by working through a case analysis on the relative order of a_j, a_k, b_j, and b_l along the y-axis. Intuitively, we can think of a_j, a_k, b_j, and b_l as the four corners of a trapezoid as in Figure 4. The diagonals of the trapezoid represent the distances under the matching μ, while the legs of the trapezoid represent the distances when we pair $\overline{a_j}$ with $\overline{b_j}$ and $\overline{a_k}$ with $\overline{b_l}$. The maximum of the lengths of the legs will always be less than the maximum of the lengths of the diagonals. Adjusting the lengths of the top and bottom bases (which amounts to changing the order of a_j, a_k, b_j, and b_l along the y-axis) does not change this fact. Therefore, matching $\overline{b_j}$ with $\overline{a_j}$ instead of $\overline{a_k}$ does not increase the cost of the matching.

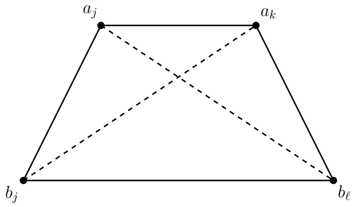

Figure 4: A trapezoid formed by a_j, a_k, b_j, and b_l.

In the second case, if $\overline{b_j}$ is matched to $\overline{0}$, there must be some $\overline{a_k}$ with $k \geq j$ that is matched to $\overline{0}$, as well. If we were to instead match $\overline{b_j}$ to $\overline{a_k}$, this does not increase the cost of the matching since $\max\{b_j, a_k\} \geq |a_k - b_j|$ (i.e., the original cost is greater than the new cost). After this rematching, $\overline{b_j}$ is no longer matched to $\overline{0}$ and this reverts to the first case. Similarly, if $\overline{a_j}$ is matched to $\overline{0}$, it may be rematched in a similar manner.

By looking at all the pairings where μ and μ^* differ (in increasing order of indices), pairing $\overline{a_i}$ with $\overline{b_i}$ instead of $\mu(\overline{a_i})$ (and similarly, pairing $\overline{b_i}$ with $\overline{a_i}$ rather than what it was paired with under μ) always results in the same or lower cost matching. Therefore, $C_{\mu^*} \leq C_\mu$ for all matchings μ; hence, $d_B(D_1, D_2) = C_{\mu^*} = \max_{i=1}^{n} |a_i - b_i|$. □

To see how this applies to the computation of the intrinsic Čech distance between two metric graphs, let G_1 be a metric graph with a shortest system of m loops of lengths $0 < 2t'_1 \leq \cdots \leq 2t'_m$, and let G_2 be a metric graph with a shortest system of n loops of lengths $0 < 2s_1 \leq \cdots \leq 2s_n$. Without loss of generality, suppose $n \geq m$. From [16], the 1-dimensional intrinsic Čech persistence diagrams of G_1 and G_2 are the multisets of points $\mathrm{Dg}_1 IC_{G_1} = \left\{\left(0, \frac{t'_1}{2}\right), \ldots, \left(0, \frac{t'_m}{2}\right)\right\}$ and $\mathrm{Dg}_1 IC_{G_2} = \left\{\left(0, \frac{s_1}{2}\right), \ldots, \left(0, \frac{s_n}{2}\right)\right\}$. In order to apply Theorem 5, we add $n-m$ copies of the point $(0,0)$ at the start of the list of points in $\mathrm{Dg}_1 IC_{G_1}$, i.e., let

$$\mathrm{Dg}_1 IC_{G_1} = \left\{\left(0, \frac{t_1}{2}\right), \ldots, \left(0, \frac{t_n}{2}\right)\right\},$$

where $t_1 = \cdots = t_{n-m} = 0$, $t_{n-m+1} = t'_1, \ldots,$ and $t_n = t'_m$.

Corollary 6. *Let G_1 and G_2 be as above. Then*

$$d_{IC}(G_1, G_2) = \max_{i=1}^{n} \frac{|s_i - t_i|}{2}.$$

Note that the corollary implies that the intrinsic Čech distance could instead be defined as a matching distance between the sets of loop lengths in shortest systems of loops for G_1 and G_2, which does not require Čech complexes or persistent homology.

4 Relating the intrinsic Čech and persistence distortion distances for a bouquet graph and an arbitrary graph

4.1 Feasible regions in persistence diagrams

Our eventual goal for our main theorem (Theorem 11) is to estimate a lower bound for the persistence distortion distance between metric graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ so that we can compare it with the intrinsic Čech distance between them, given in Corollary 6. A fundamental part of this process relies on the notion of a *feasible region* for a point in a given persistence diagram lying on the y-axis.

Definition 7. *The **feasible region** for a point $\overline{s} := (0, s) \in \mathbb{R}^2$ is defined as*

$$F_{\overline{s}} = \{z = (z_1, z_2) : 0 \leq z_1 \leq z_2, s \leq z_2 \leq z_1 + s\}.$$

An illustration of a feasible region is shown in Figure 5.

The following lemma establishes an important property of feasible regions that will be used later in the proof of the main theorem.

Lemma 8. *Given any point $z \in F_{\overline{s}}$ and any point $\overline{t} = (0, t)$, $||\overline{s} - \overline{t}||_1 \leq ||z - \overline{t}||_1$.*

Proof. We proceed with a simple case analysis using the definition of $F_{\overline{s}}$. Let $z = (z_1, z_2)$.

Case 1: Assume $s \geq t$ so that $||\overline{s} - \overline{t}||_1 = s - t$. By the definition of $F_{\overline{s}}$, we have $z_2 \geq s$ and thus

$$||z - \overline{t}||_1 = z_1 + z_2 - t \geq z_1 + s - t \geq s - t = ||\overline{s} - \overline{t}||_1.$$

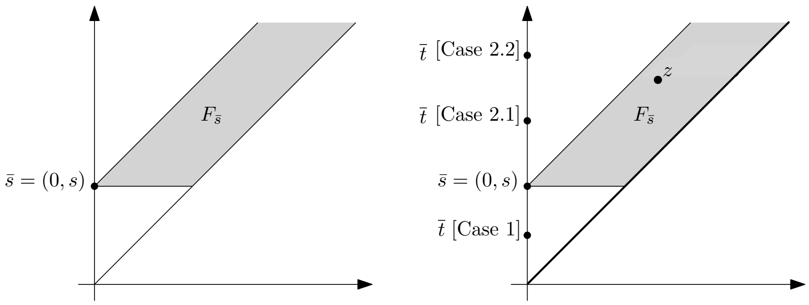

Figure 5: Left: An illustration of the feasible region for $\overline{s}$. Right: The three cases within the proof of Lemma 8.

Case 2.1: If $s < t$, then $||\overline{s} - \overline{t}||_1 = t - s$. If $t \leq z_2$, then since $z_1 \geq z_2 - s$ and $z_2 \geq s$,

$$||z - \overline{t}||_1 = z_1 + z_2 - t \geq (z_2 - s) + z_2 - t \geq t - s + t - t = t - s = ||\overline{s} - \overline{t}||_1.$$

Case 2.2: If $s < t$ but $t > z_2$, then since $z_2 \leq z_1 + s$, it follows that

$$||z - \overline{t}||_1 = z_1 + t - z_2 \geq z_1 + t - (z_1 + s) = t - s = ||\overline{s} - \overline{t}||_1.$$

The lemma now follows. □

4.2 Properties of the geodesic distance function for an arbitrary metric graph

Let $G = (V, E)$ be an arbitrary metric graph with shortest system of loops of lengths $2s_1, \cdots, 2s_n$. Fix an arbitrary base point $v \in |G|$ and consider $\mathrm{Dg}(f_v)$, as defined in Section 2.2. Let T_v denote the shortest path tree in G rooted at v.

We consider the base point $v \in |G|$ to be a graph node of G; that is, we add it to V if necessary. We further assume that the pair (G, v) is "generic" in the sense that there do not exist two or more shortest paths from the base point v to any graph node of G in V. This will be important later to ensure that local maxima of f_v are not at graph nodes but instead occur at internal points within edges of G. Later we will assume that (G, v) is generic for all $v \in |G|$ since for any input metric graph G and base point v, we can perturb it to be one that is generic within arbitrarily small Gromov-Hausdorff distance. In particular, Figure 3 is not generic, but due to the stability of both the persistence distortion and intrinsic Čech distances with respect to the Gromov-Hausdorff distance [11, 13], there exists a small perturbation to ensure genericity.

For simplicity, when v is fixed, we shall omit v in our notation and speak of the persistence diagram $D := \mathrm{Dg}(f_v)$, the function $f := f_v$, and the shortest path tree $T := T_v$.

We present three straightforward observations, the first of which follows immediately from the definition of the shortest path tree and the Extreme Value Theorem.

Observation 1. *The shortest path tree T of G has $|V| - 1$ edges, and there are $|E| - |V| + 1$ non-tree edges. For each non-tree edge $e \in E \setminus T$, there exists a unique $u \in e$ such that $f(u)$ is a local maximum value of f.*

Figure 6 contains an example of a metric graph on the left and the corresponding shortest path tree illustrating Observation 1 in the middle.

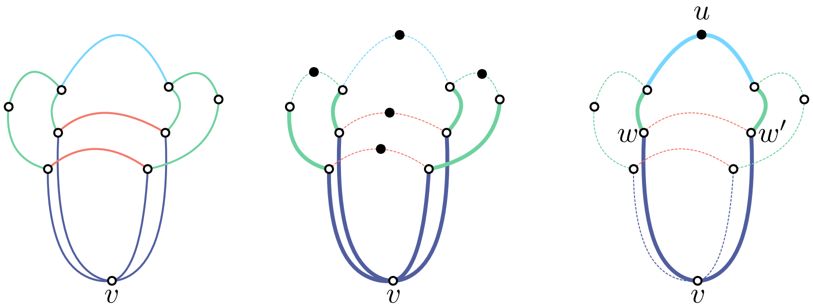

Figure 6: Left: An example of a metric graph G. The dark blue, light blue, red, and green edges are of lengths 4, 3, 2, and 1, respectively, and v is the base point. Middle: An illustration of Observation 1 for the shortest path tree T of G. Thick blue and green edges are part of the shortest path tree T, and the others are the non-tree edges. Black points represent local maxima of f. Right: An illustration of Observation 3. Thick edges are part of γ, $f(v) = p$, and p' is either $f(w)$ or $f(w')$ depending on the relative positions of w and w'.

Note that every feature in the persistence diagram D must die at a point in the graph that is an up-fork, i.e., a point coupled with a pair of adjacent directions along which the function f is increasing. Since there are no local minimum points of f (except for v itself), these must be vertices in the graph of degree at least 3 (see, e.g., [21]).

Observation 2. *The diagram D is a multiset of points $\{(p_i, f(u_i)) \mid i \in \{1, \ldots, |E| - |V| + 1\}\}$, where for every i, $p_i = f(w)$ for some graph node $w \in V$.*

The final observation relates to points belonging to cycles in G that yield local maximum values of f (see [2]).

Observation 3. *Let γ be an arbitrary cycle in G. If u is the point in γ corresponding to the largest local maximum value of f, let p be the lowest function value of f for all points in γ. Then there is a point in the persistence diagram D of the form $(p', f(u))$, where $p' \geq p$.*

See Figure 6 (right) for an illustration.

To delve further into this, let $\{\gamma_1, \ldots, \gamma_n\}$ denote the elements of the shortest system of loops for G listed in non-decreasing order of loop length, $s_1 \leq \ldots \leq s_n$.

Lemma 9. *Consider a cycle $\gamma = \gamma_{i_1} + \ldots + \gamma_{i_m}$ $(m \leq n)$, where each γ_{i_k} $(1 \leq k \leq m)$ is an element of the shortest system of loops for G and $i_1 \leq i_2 \leq \ldots \leq i_m$. Suppose the edge $e \in \gamma$ contains the point u in γ with the largest local maximum value of f. Then $f(u) \geq s_{i_m}$, where s_{i_m} is half the length of cycle γ_{i_m} in the shortest system of loops.*

Proof. The proof of this lemma will proceed by contradiction. We assume instead that $f(u) < s_{i_m}$. Under this assumption we will show that the cycle γ can be decomposed as the sum of cycles $\{c_j\}$, each of which has length strictly smaller than s_{i_m}, which contradicts the fact that $\{\gamma_{i_j}\}$ are in the shortest system of loops and implies that $f(u)$ must be at least s_{i_m}.

Since each γ_{i_k} $(1 \leq k \leq m)$ is an element of the shortest system of loops for G and $i_1 \leq i_2 \leq \ldots \leq i_m$, this implies that $s_{i_1} \leq \cdots \leq s_{i_m}$, where $2s_{i_k}$ is the length of cycle γ_{i_k} in the shortest system of loops of G.

The cycle γ in G must contain at least one non-tree edge as it is a cycle. Let $e_1, \ldots, e_\ell = e$ be all non-tree edges of G with largest function value at most $f(u)$. Assume they contain maximum points $u_1, \ldots, u_\ell = u$, respectively, where the edges and maxima are sorted in order of increasing function value of f.

For two points $x, y \in |T|$, let $\alpha(x, y)$ denote the unique tree path from x to y within the shortest path tree. For each $j \in \{1, \ldots, \ell\}$, let $e_j = (e_j^0, e_j^1)$ and let c_j denote the cycle $c_j = \alpha(v, e_j^1) \circ e_j \circ \alpha(e_j^0, v)$. By assumption, since $u = u_\ell$ is the point in γ with the largest local maximum value of f and $f(u) < s_{i_m}$, it follows that the length of every cycle c_j is less than s_{i_m}. However, the set of cycles $\{c_1, \ldots, c_\ell\}$ forms a basis for the subgraph of G spanned by all edges containing only points of function value at most $f(u)$. Therefore, we may represent γ as a linear combination of cycles from the set $\{c_1, \ldots, c_\ell\}$, i.e., γ may be decomposed into shorter cycles, each of length less than $s_{i_m} = \dfrac{\mathit{length}(\gamma_{i_m})}{2}$. This is a contradiction to the fact that $\gamma_{i_1}, \ldots, \gamma_{i_m}$ are elements of the shortest system of loops for G. Hence, we conclude that $f(u) \geq s_{i_m}$. □

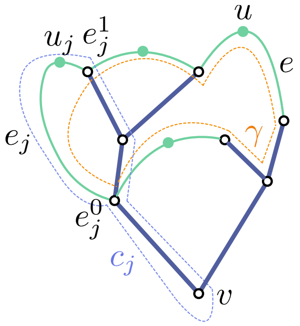

Figure 7: An illustration of the proof of Lemma 9. In this case, γ is the sum of three smaller cycles and there are four non-tree edges highlighted in green. One c_j is shown corresponding to the local maximum u_j.

An example that illustrates the proof of Lemma 9 is shown in Figure 7. Later we will use the following simpler version of Lemma 9, where γ is a single element of the shortest system of loops.

Corollary 10. *Let γ be an element of the shortest system of loops for G with a length $2s$, and let u denote the point in any edge of γ with largest maximum value of f. Then $f(u) \geq s$.*

4.3 The main theorem and its proof

We are now ready to establish a comparison of the intrinsic Čech and persistence distortion distances between a bouquet metric graph and an arbitrary metric graph.

Theorem 11. *Let G_1 and G_2 be finite metric graphs such that G_1 is a bouquet graph and G_2 is arbitrary. Then*

$$d_{IC}(G_1, G_2) \leq \frac{1}{2} d_{PD}(G_1, G_2).$$

Proof. Let G_1 be a bouquet graph consisting of m cycles of lengths $0 < 2t'_1 \leq \ldots \leq 2t'_m$, all sharing one common point $o \in |G_1|$. Let G_2 be an arbitrary metric graph with shortest system of loops consisting of n loops of lengths $2s_1, \cdots, 2s_n$ listed in non-decreasing order. In what follows, we suppose $n \geq m$; the case when $m \geq n$ proceeds similarly. As before, we obtain a sequence of length n, $2t_1 \leq 2t_2 \cdots \leq 2t_n$ (where $t_1 = \cdots = t_{n-m} = 0$, $t_{n-m+1} = t'_1, \cdots$, and $t_n = t'_m$). Let f and g denote the geodesic distance functions on G_1 and G_2, respectively.

First, as in Corollary 6, the intrinsic Čech distance between G_1 and G_2, denoted by δ, is

$$\delta := d_{IC}(G_1, G_2) = \max_{i=1}^{n} \frac{|s_i - t_i|}{2}. \tag{1}$$

Second, note that the persistence diagram $D_1 := \mathrm{Dg}(f_o)$ with respect to the base point o is $D_1 = \{(0, t_1), \cdots, (0, t_n)\}$ (of course, this may include some copies of $(0, 0)$ if $m < n$). Next, fix an arbitrary base point $v \in |G_2|$ and consider the persistence diagram $D_2 := \mathrm{Dg}(g_v)$. Consider the abstract persistence diagram $D^\star := \{(0, s_1), \cdots, (0, s_n)\} = \{\overline{s_1}, \ldots, \overline{s_n}\}$ that consists only of points on the y-axis at the s_i values. Unless G_2 is also a bouquet graph, $D^\star$ is not necessarily in $\Phi(|G_2|)$. Nevertheless, we will use this persistence diagram as a point of comparison and relate points in D_2 to $D^\star$. Notice that a consequence of Theorem 5 is that

$$d_B(D_1, D^\star) = \max_{i=1}^{n} |s_i - t_i| = 2\delta. \tag{2}$$

In order to accomplish our objective of relating points in D_2 with points in the ideal diagram $D^\star$, we need the following lemma relating to feasible regions, which were introduced in Section 4.1.

Lemma 12. *Let $D' = \{z_1, \ldots, z_n\}$ be an arbitrary persistence diagram such that $z_i \in F_{\overline{s_i}}$. Then $d_B(D_1, D^\star) \leq d_B(D_1, D')$.*

Proof. Consider the optimal bottleneck matching between D_1 and D'. According to Lemma 8, if the point $\overline{t_j} = (0, t_j) \in D_1$ is matched to $z_i \in D'$ under this optimal matching, the matching of $\overline{s_i} = (0, s_i) \in D^\star$ to $\overline{t_j}$ will yield a smaller distance. In other words, the induced bottleneck matching between D_1 and $D^\star$, which is equal to 2δ, can only be smaller than $d_B(D_1, D')$. □

The outline of the remainder of the proof of Theorem 11 is as follows. Theorem 13 shows that one can assign points in D_2 to the points in $D^\star$ in such a way that the condition in

Lemma 12 is satisfied. The fact that one can assign points in the fixed persistence diagram D_2 to the distinct feasible regions $F_{\overline{s_i}}$ relies on the series of structural observations and results in Section 4.2, along with an application of Hall's marriage theorem. Finally, the inequality in Lemma 12 and the definition of the persistence distortion distance imply that

$$2\delta = d_B(D_1, D^\star) \leq \inf_{v \in |G_2|} d_B(D_1, D_2) \leq d_{PD}(G_1, G_2), \tag{3}$$

which, together with (1), completes the proof of Theorem 11. Notice that if any (G_2, v) is not generic, we can perturb either G_2 or v so that $|d_B(D_1, D_2) - d_B(D_1, D_2')| < \epsilon$ for any $\epsilon > 0$.

The following theorem establishes the existence of a one-to-one correspondence between points in $D^\star$ and points in D_2. The goal is to construct a bipartite graph $\widehat{G} = (D^\star, D_2, \widehat{E})$, where there is an edge $\hat{e} \in \widehat{E}$ from $\overline{s_i} \in D^\star$ to $z \in D_2$ if and only if $z \in F_{\overline{s_i}}$. To prove the theorem, we invoke Hall's marriage theorem, which requires showing that for any subset S of points in $D^\star$, the number of neighbors of S in D_2 is at least $|S|$.

Theorem 13. *The graph $\widehat{G}$ contains a perfect matching.*

Proof. For simplicity, let $T = T_v$ and $g = g_v$. First, note that there is a one-to-one correspondence $\Psi : E_2 \setminus T \to D_2$ between the set of non-tree edges in G_2 (each of which contains a unique maximum point of g) and the set of points in D_2. In particular, from Observations 1 and 2, the birth-time of each point in D_2 uniquely corresponds to a local maximum u_e within a non-tree edge e of G_2. The assumption that v is generic implies that the local maximum u_e occurs within an edge and not at a vertex.

Fix an arbitrary subset $S \subseteq D^\star$ with $|S| = a$. In order to apply Hall's marriage theorem, we must show that there are at least a neighbors of S in $\widehat{G}$. We achieve this via an iterative procedure which we now describe. The procedure begins at step $k = 0$ and will end after a iterations. Elements in $S = \{\overline{s_{i_1}}, \ldots, \overline{s_{i_a}}\}$ are processed in non-decreasing order of their values, which also means that $i_1 < i_2 < \cdots < i_a$. At the start of the k-th iteration, we will have processed the first k elements of S, denoted $S_k = \{\overline{s_{i_1}}, \ldots, \overline{s_{i_k}}\}$, where for each $\bar{s} := \overline{s_{i_h}} \in S_k$ that we have processed $(1 \leq h \leq k)$, we have maintained the following three invariances:

Invariance 1: $\bar{s}$ is associated to a unique edge $e_{\bar{s}} \in E_2 \setminus T$ containing a unique maximum $u_{e_{\bar{s}}}$ such that $\Psi(e_{\bar{s}}) \in D_2$ is a neighbor of $\bar{s}$. We say that $e_{\bar{s}}$ and $u_{e_{\bar{s}}}$ are *marked by* $\bar{s}$.

Invariance 2: : $\bar{s}$ is also associated to a cycle $\widetilde{\gamma}_h = \gamma_{i_h} + \sum \gamma_\ell$ (where the sum ranges over all ℓ belonging to some index set $J_h \subset \{1, \ldots, i_h - 1\}$), such that $e_{\bar{s}}$ contains the point in $\widetilde{\gamma}_h$ with the largest value of g.

Invariance 3: : $height(\widetilde{\gamma}_h) \leq s_{i_h}$, where $height(\gamma) = \max_{x \in \gamma} g(x) - \min_{x \in \gamma} g(x)$ represents the height (i.e., the maximal difference in the g function values) of a given loop γ.

Set $\overline{S_k} = S \setminus S_k = \{\overline{s_{i_{k+1}}}, \ldots, \overline{s_{i_a}}\}$, denoting the remaining elements from S to be processed. Our goal is to identify a new neighbor in D_2 for element $\overline{s_{i_{k+1}}}$ from $\overline{S_k}$ satisfying the three

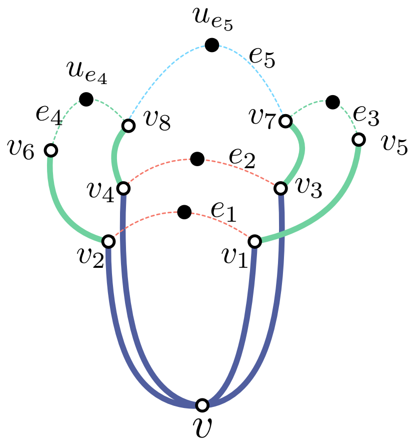

Figure 8: One shortest system of loops for this graph $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_5\}$ consists of the following: $\gamma_1 = v_4v_3v_7v_8v_4$, $\gamma_2 = v_2v_1v_5v_7v_8v_6v_2$, $\gamma_3 = vv_1v_2v$, $\gamma_4 = vv_3v_4v$, and $\gamma_5 = vv_1v_5v_7v_3v$. Their lengths are $7, 9, 10, 10$, and 11, respectively, generating points $\overline{s_1} = (0, 3.5), \overline{s_2} = (0, 4.5), \overline{s_3} = (0, 5), \overline{s_4} = (0, 5)$, and $\overline{s_5} = (0, 5.5)$. Now, imagine we have $S = \{\overline{s_1}, \overline{s_2}\}$. When we process $\overline{s_1}$, we consider the shortest loop γ_1 corresponding to it, and edge e_5 contains the maximum u_{e_5} of highest g function value among points in γ_1. Since this edge is not yet marked, we set $\widetilde{\gamma}_1 = \gamma_1$, and mark e_5 by $\overline{s_1}$. Next, we consider $\overline{s_2}$. We start with the highest edge in the shortest loop γ_2, which is again e_5. However, at this point, e_5 is already marked with $\overline{s_1}$, we thus instead take the combination of γ_2 with the loop associated to e_5, that is, we set $\widehat{\gamma}_2 = \gamma_2 + \widetilde{\gamma}_1 = \gamma_2 + \gamma_1 = v_1v_5v_7v_3v_4v_8v_6v_2v_1$. The edge containing the highest function value in this loop is e_4 (assume we break the tie of e_4 and e_3 by using their indices). Since e_4 is not yet marked, we set $\widetilde{\gamma}_2 = \widehat{\gamma}_2$ and mark e_4 with this.

invariances. Once we have done so, we will then set $S_{k+1} = S_k \cup \{s_{i_{k+1}}\}$ and move on to the next iteration in the procedure.

Note that $\overline{s_{i_{k+1}}}$ corresponds to an element $\gamma_{i_{k+1}}$ of the shortest system of loops for G_2. Let e be the edge in $\gamma_{i_{k+1}}$ containing the maximum u_e of highest g function value among all edges in $\gamma_{i_{k+1}}$. There are now two possible cases to consider, and we will demonstrate how to obtain a new neighbor for $\overline{s_{i_{k+1}}}$ in either case.

In the first case, suppose u_e is not yet marked by a previous element in S. In this case, $e_{\overline{s_{i_{k+1}}}} = e$ and $\widetilde{\gamma}_{i_{k+1}} = \gamma_{i_{k+1}}$. We claim that the point $(p_e, g(u_e))$ in the persistence diagram D_2 corresponding to the maximum u_e is contained in the feasible region $F_{\overline{s_{i_{k+1}}}}$. In other words, $s_{i_{k+1}} \leq g(u_e) \leq p_e + s_{i_{k+1}}$. Indeed, by Lemma 9, $s_{i_{k+1}} \leq g(u_e)$, and by Observation 3,

$$g(u_e) - s_{i_{k+1}} \leq lowest(\gamma_{i_{k+1}}) \leq p_e,$$

where $lowest(\gamma_{i_{k+1}}) := \min_{x \in \gamma_{i_{k+1}}} g(x)$. Thus, $(p_e, g(u_e)) \in D_2$ is a new neighbor for $\overline{s_{i_{k+1}}} \in S$ since it is contained in $F_{\overline{s_{i_{k+1}}}}$. Consequently, we mark e and u_e by $\overline{s_{i_{k+1}}}$ and continue with the next iteration.

In the second case, the maximum point u_e has already been marked by a previous element $s_{j_1} \in S_k$ and been associated to a cycle $\widetilde{\gamma}_{j_1}$. Observe that $s_{j_1} \leq s_{i_{k+1}}$ since our procedure processes elements of S in non-decreasing order of their values (and thus $j_1 < i_{k+1}$). We must now identify an edge other than e for $s_{i_{k+1}}$ satisfying the three invariance properties. To this end, let $\widehat{\gamma}_1 = \gamma_{i_{k+1}} + \widetilde{\gamma}_{j_1}$, and let e_1 be the edge containing the maximum in $\widehat{\gamma}_1$ with largest function value. If e_1 is unmarked, we set $e_{\overline{s_{i_{k+1}}}} = e_1$. Otherwise, if e_1 is marked by some cycle γ_{j_2}, we construct the loop $\widehat{\gamma}_2 = \widehat{\gamma}_1 + \widetilde{\gamma}_{j_2} = \gamma_{i_{k+1}} + \widetilde{\gamma}_{j_1} + \widetilde{\gamma}_{j_2}$ with the purpose of erasing the already-marked edge e_1 from the loop and finding a new edge e_2 containing the maximum function value of $\widehat{\gamma}_2$. We continue this process until we find $\widehat{\gamma}_\eta = \gamma_{i_{k+1}} + \widetilde{\gamma}_{j_1} + \widetilde{\gamma}_{j_2} + \ldots + \widetilde{\gamma}_{j_\eta}$ such that the edge e_η containing the point of maximum function value of $\widehat{\gamma}_\eta$ is not marked. Once we arrive at this point, we set $\widetilde{\gamma}_{i_{k+1}} = \widehat{\gamma}_\eta$ and $e_{\overline{s_{i_{k+1}}}} = e_\eta$, so that the edge e_η and corresponding maximum u_{e_η} are marked by $\overline{s_{i_{k+1}}}$. For a concrete example of this procedure, see Figure 8 and its caption.

The reason that the procedure outlined above must indeed terminate is as follows. Each time a new $\widetilde{\gamma}_{j_\nu}$ is added to a cycle $\widehat{\gamma}_{j_{\nu-1}}$ (for $\nu \in \{1, \ldots, \eta\}$), it is because the edge containing the maximum point of $\widehat{\gamma}_{j_{\nu-1}}$ with largest function value is marked by $\overline{s_{j_\nu}}$. Note that $j_\nu \neq j_\beta$ for $\nu \neq \beta$ (as during the procedure, the edge e_i containing the maximum function value in the cycle $\widehat{\gamma}_i$ are all distinct), each $j_\nu < i_{k+1}$, and $\overline{s_{j_\nu}} \in S_k$. Furthermore, Invariance 2 guarantees that $\widehat{\gamma}_\eta$ cannot be empty, as each cycle $\widetilde{\gamma}_{j_\nu}$ can be written as a linear combination of elements in the shortest system of loops with indices at most j_ν. As $j_\nu < i_{k+1}$, the cycle $\gamma' = \widetilde{\gamma}_{j_1} + \widetilde{\gamma}_{j_2} + \ldots + \widetilde{\gamma}_{j_\eta}$ can be represented as a linear combination of basis cycles with indices strictly smaller than i_{k+1}. In other words, $\gamma_{i_{k+1}}$ and γ' must be linearly independent, and thus $\widehat{\gamma}_\eta = \gamma_{i_{k+1}} + \gamma'$ cannot be empty. Again, $j_\nu \neq j_\beta$ for $\nu \neq \beta$ and each $j_\nu < i_{k+1}$, and thus it follows that after at most k iterations, we will obtain a cycle whose highest valued maximum and corresponding edge are not yet marked.

Now, we must show that the three invariances are satisfied as a result of the process described in this second case. To begin, we point out that Invariance 2 holds by construction. Next, the following lemma establishes Invariance 3.

Lemma 14. *For $\widetilde{\gamma}_{i_{k+1}} = \widehat{\gamma}_\eta = \gamma_{i_{k+1}} + \widetilde{\gamma}_{j_1} + \widetilde{\gamma}_{j_2} + \ldots + \widetilde{\gamma}_{j_\eta}$ as above, height$(\widetilde{\gamma}_{i_{k+1}}) \leq s_{i_{k+1}}$.*

Proof. Set $\widehat{\gamma}_0 = \gamma_{i_{k+1}}$, and for $\nu \in \{1, \ldots, \eta\}$, set $\widehat{\gamma}_\nu = \gamma_{i_{k+1}} + \widetilde{\gamma}_{j_1} + \cdots + \widetilde{\gamma}_{j_\nu}$. Using induction, we will show that $height(\widehat{\gamma}_\nu) \leq s_{i_{k+1}}$ for any $\nu \in \{0, \ldots, \eta\}$. The inequality obviously holds for $\nu = 0$. Suppose it holds for all $\nu \leq \rho < \eta$, and consider $\nu = \rho + 1$ where $\widehat{\gamma}_{\rho+1} = \widehat{\gamma}_\rho + \widetilde{\gamma}_{j_{\rho+1}}$. The cycle $\widetilde{\gamma}_{j_{\rho+1}}$ is added as the edge e_ρ of $\widehat{\gamma}_\rho$ containing the current maximum point of highest value of g has already been marked by $\overline{s_{j_{\rho+1}}}$ with $j_{\rho+1} < i_{k+1}$. By Invariance 2, e_ρ must also be the edge in $\widetilde{\gamma}_{j_{\rho+1}}$ containing the point of maximum g function value, which we denote by $g(e_\rho)$. Therefore, after the addition of $\widehat{\gamma}_\rho$ and $\widetilde{\gamma}_{j_{\rho+1}}$,

$$\begin{aligned} &\text{(i) } highest(\widehat{\gamma}_{\rho+1}) := \max_{x \in \widehat{\gamma}_{\rho+1}} g(x) \leq g_v(e_\rho), \text{ and} \qquad (4) \\ &\text{(ii) } lowest(\widehat{\gamma}_{\rho+1}) := \min_{x \in \widehat{\gamma}_{\rho+1}} g(x) \geq \min\{\ lowest(\widehat{\gamma}_\rho),\ lowest(\widetilde{\gamma}_{j_{\rho+1}})\ \}. \end{aligned}$$

By the induction hypothesis, $height(\widehat{\gamma}_\rho) \leq s_{i_{k+1}}$, while by Invariance 3, $height(\widetilde{\gamma}_{j_{\rho+1}}) \leq s_{j_{\rho+1}} \leq s_{i_{k+1}}$. By (ii) of equation (4), it then follows that

$$lowest(\widehat{\gamma}_{\rho+1}) \geq \min\{g(e_\rho) - height(\widehat{\gamma}_\rho), g(e_\rho) - height(\widetilde{\gamma}_{j_{\rho+1}})\} \geq g(e_\rho) - s_{i_{k+1}}.$$

Combining this with (i) of equation (4), we have that $height(\widehat{\gamma}_{\rho+1}) \leq s_{i_{k+1}}$. The lemma then follows by induction. □

Finally, we show that Invariance 1 also holds. Since $\widetilde{\gamma}_{i_{k+1}} = \widehat{\gamma}_\eta = \gamma_{i_{k+1}} + \gamma'$, with γ' defined as above, by Lemma 9, we have that $g(u_{e_\eta}) \geq s_{i_{k+1}}$. Suppose u_{e_η} is paired with some graph node w so that $p_{e_\eta} = g(w)$. As the height of $\widetilde{\gamma}_{i_{k+1}}$ is at most $s_{i_{k+1}}$ (Lemma 14), combined with Observation 3, we have that

$$g(u_{e_\eta}) - s_{i_{k+1}} \leq lowest(\widetilde{\gamma}_{i_{k+1}}) \leq p_{e_\eta}.$$

This implies that the point $(p_{e_\eta}, g(u_{e_\eta})) \in F_{\overline{s_{i_{k+1}}}}$, establishing Invariance 1.

We continue the process described above until $k = a$. At each iteration, when we process $\overline{s_{i_k}}$, we add a new neighbor for elements in S. In the end, after processing all of the a elements in S, we find a neighbors for S, and the total number of neighbors in $\widehat{G}$ of elements in S can only be larger. Since this holds for any subset S of $D^\star$, the condition for Hall's theorem is satisfied for the bipartite graph $\widehat{G}$. This implies that there exists a perfect matching in $\widehat{G}$, completing the proof of Theorem 13. □

Theorem 11 now follows from Lemma 12 and equation (1). □

4.4 Proving the conjecture when both graphs are trees of loops

The techniques of Theorem 11 are specific to the case when one of the graphs is a bouquet graph. However, we can prove Conjecture 1 in another setting, as well: the case when both graphs are *trees of loops*.

Definition 15. *A* ***tree of loops*** *is a metric graph constructed via wedge sums of cycles and edges.*

See Figure 9 for an example. The fact that the inequality holds when both graphs are trees of loops follows from an application of the following lemma.

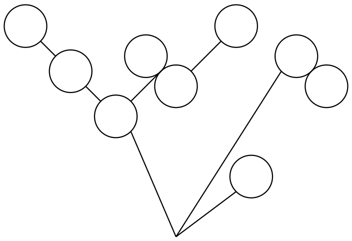

Figure 9: An example of a tree of loops.

Lemma 16. *Let P and Q be two persistence diagrams with finite numbers of off-diagonal points. Let d_1 and d_2 be distances defined between points in P and points in Q such that $d_1(p, q) \leq d_2(p, q)$ for every $p \in P, q \in Q$. Then $d_B(P, Q)$ under distance d_1 is less than or equal to $d_B(P, Q)$ under distance d_2.*

Proof. The bottleneck distance under a particular distance d is given by

$$d_B(P,Q) = \min_{\mu} \max_{p} d(p, \mu(p)),$$

where the minimum is taken over all matchings $\mu : P \to Q$. If we consider a fixed matching μ, the relationship between d_1 and d_2 implies that

$$\max_p d_1(p, \mu(p)) \leq \max_p d_2(p, \mu(p)). \tag{5}$$

Let μ' be the matching that achieves the minimum for distance d_2. The inequality (5) together with this minimum implies that

$$d_B(P,Q) \text{ under } d_1 = \min_{\mu} \max_{p} d_1(p,\mu(p)) \leq \max_p d_1(p,\mu'(p)) \leq \max_p d_2(p,\mu'(p)) = d_B(P,Q) \text{ under } d_2.$$

□

Proposition 17. *Let G_1 and G_2 be two finite metric graphs such that both are trees of loops. Then*

$$d_{IC}(G_1, G_2) \leq \frac{1}{2} d_{PD}(G_1, G_2).$$

Proof. Let G_1 and G_2 be trees of loops of lengths $2t'_1, \ldots, 2t'_m$ and $2s_1, \ldots, 2s_n$, respectively, each set listed in non-decreasing order. Without loss of generality, suppose $n \geq m$. First, as in Corollary 6, the intrinsic Čech distance between G_1 and G_2 is $d_{IC}(G_1, G_2) = \max_{i=1}^{n} \dfrac{|s_i - t_i|}{2}$ where $t_1 = \cdots = t_{n-m} = 0$, $t_{n-m+1} = t'_1, \ldots,$ and $t_n = t'_m$. Suppose $d_{IC}(G_1, G_2) = \dfrac{|t_k - s_k|}{2}$ for some k, $1 \leq k \leq n$. Let f and g denote the geodesic distance functions on G_1 and G_2, respectively.

For trees of loops, the persistence diagrams take the form $\mathrm{Dg}(f_v) = \{(p_i, p_i + t_i)\}_{1 \leq i \leq n}$ and $\mathrm{Dg}(g_w) = \{(q_i, q_i + s_i)\}_{1 \leq i \leq n}$ for $v \in G_1$ and $w \in G_2$. Here, p_i (resp., q_i) is the length of a shortest path from v (resp., w) to the closest point on the corresponding loop of length $2t_i$ (resp, $2s_i$). The proposition holds if, for any pair of persistence diagrams $\mathrm{Dg}(f_v)$ and $\mathrm{Dg}(g_w)$, $d_B(\mathrm{Dg}(f_v), \mathrm{Dg}(g_w)) \geq |t_k - s_k|$.

We will prove this by applying Lemma 16. For any $i, j \in \{1, \ldots, n\}$, let $d_1((p_i, p_i + t_i), (q_j, q_j + s_j)) = |t_i - s_j|$ and $d_2((p_i, p_i + t_i), (q_j, q_j + s_j)) = ||(p_i, p_i + t_i) - (q_j, q_j + s_j)||_1$. It's easy to show that $d_1 \leq d_2$ for every pair of points, so that the conditions of the lemma are satisfied. Notice that distance d_1 is equivalent to the case where all $p_i = q_i = 0$, i.e., the points are along the y-axis. By Theorem 5, the bottleneck distance under d_1 equals $|t_k - s_k| = 2d_{IC}(G_1, G_2)$. Therefore, the bottleneck distance between $\mathrm{Dg}(f_v)$ and $\mathrm{Dg}(g_w)$ under d_2 is at least $|t_k - s_k|$, as desired. □

5 Discussion and future work

In this paper, we compare the discriminative capabilities of the intrinsic Čech and persistence distortion distances, which are based on topological signatures of metric graphs. The intrinsic Čech signature arises from the intrinsic Čech filtration of a metric graph, and the persistence distortion signature is based on the set of persistence diagrams arising from sublevel set filtrations of geodesic distance functions from all base points in a given metric graph. A map from a metric graph to these topological signatures is "lossy" as it is not injective: two different metric graphs may map to the same signature. However, topological signatures capture structural information of graphs and understanding the relationship between the intrinsic Čech and persistence distortion distances enables us to better understand the discriminative powers of such signatures.

We conjecture that the intrinsic Čech distance is less discriminative than the persistence distortion distance for general metric graphs G_1 and G_2, so that there exists a constant $c > 0$ with $d_{IC}(G_1, G_2) \leq c \cdot d_{PD}(G_1, G_2)$. This statement is trivially true in the case when both graphs are trees as the intrinsic Čech distance is 0 while the persistence distortion distance is not. We prove a version of the conjectured inequality in the case when one of the graphs is a bouquet graph and the other is arbitrary, as well as in the case when both graphs are obtained via wedges of cycles and edges. The methods of proof in Theorem 11 and Proposition 17 rely on explicitly knowing the forms of the persistence diagrams for the geodesic distance function in the case of a bouquet graph or a tree of loops. Therefore, these methods do not readily carry over to the most general setting for arbitrary metric graphs. Nevertheless, we believe that the relationship between the intrinsic Čech and persistence distortion distances should hold for arbitrary finite metric graphs. Intuitively, the intrinsic Čech signature only captures the sizes of the shortest loops in a metric graph, whereas the persistence distortion signature takes into consideration the relative positions of such loops and their interactions with one another. We believe that proving our conjecture in greater generality is highly nontrivial and leave it as an open question.

As one example application relating the intrinsic Čech and persistence distortion signatures, the work of Pirashvili et al. [22] considers how the topological structures of chemical compounds relate to solubility in water, which is of fundamental importance in modern drug discovery. Analysis with the topological tool mapper [24] reveals that compounds with a smaller number of cycles are more soluble. The number of cycles, as well as cycle lengths, is naturally encoded in the intrinsic Čech signature. In addition, these authors also use a discrete persistence distortion signature – where only the graph nodes, i.e., the atoms, serve as base points – to show that nearby compounds have similar levels of solubility. Although we conjecture that the intrinsic Čech distance is less discriminative than the persistence distortion distance, it might be sufficient in this particular analysis since solubility is highly correlated with the number of cycles of a chemical compound, that is, with the intrinsic Čech signature [16]. It would be interesting to investigate other applications of the intrinsic Čech and persistence distortion signatures in the context of data modeled by metric graphs.

In addition, recall from the definition of the persistence distortion distance the map $\Phi : |G| \to SpDg$, $\Phi(v) = \mathrm{Dg}(f_v)$. The map Φ is interesting in its own right. For instance, what can be said about the set $\Phi(|G|)$ in the space of persistence diagrams for a given G? Given only the set $\Phi(|G|) \subset SpDg$, what information can one recover about the graph G? Oudot and Solomon [21] show that there is a dense subset of metric graphs (in the Gromov–Hausdorff topology, and indeed an open dense set in the so-called fibered topology) on which their barcode transform via the map Φ is globally injective up to isometry. They also prove its local injectivity on the space of metric graphs. Another question of interest is, how does the map Φ induce a stratification in the space of persistence diagrams? Finally, it would also be worthwhile to compare the discriminative capacities of the persistence distortion and intrinsic Čech distances to other graph distances, such as the interleaving and functional distortion distances in the special case of Reeb graphs.

6 Acknowledgements

The authors are grateful for the Women in Computational Topology (WinCompTop) workshop at the Institute for Mathematics and its Applications (IMA) for initiating our research collaboration. Participant travel support to this workshop was made possible through the National Science Foundation grant DMS-1619908. The IMA also hosted this collaborative group at a follow-up workshop. The American Institute of Mathematics (AIM) has supported our group through the Structured Quartet Research Ensembles (SQuaRE) program. EP was supported by the Pacific Northwest National Laboratory (PNNL). PNNL is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. RS was partially supported by the Simons Collaboration Grant 318086 in the early stages of this project and NSF DMS-1854705. BW was partially supported by the National Science Foundation under grants IIS-1513616 and DBI-1661375. YW was partially supported by the National Science Foundation under grants CCF-1740761, DMS-1547357, and RI-1815697. LZ was partially supported by the National Science Foundation under grant CDS&E-MSS-1854703. The authors thank the anonymous reviewers for their helpful feedback in clarifying and improving the manuscript.

References

[1] Aanjaneya, M., Chazal, F., Chen, D., Glisse, M., Guibas, L., Morozov, D.: Metric graph reconstruction from noisy data. International Journal of Computational Geometry & Applications **22**(04), 305–325 (2012)

[2] Agarwal, P.K., Edelsbrunner, H., Harer, J., Wang, Y.: Extreme elevation on a 2-manifold. Discrete & Computational Geometry **36**(4), 553–572 (2006)

[3] Babai, L.: Graph isomorphism in quasipolynomial time. ACM Symposium on Theory of Computing pp. 684–697 (2016)

[4] Bauer, U., Ge, X., Wang, Y.: Measuring distance between Reeb graphs. Symposium on Computational Geometry (2014)

[5] Bauer, U., Munch, E., Wang, Y.: Strong equivalence of the interleaving and functional distortion metrics for Reeb graphs. Symposium on Computational Geometry **34**, 461–475 (2015)

[6] Burago, D., Burago, Y., Ivanov, S.: A course in metric geometry, vol. 33. American Mathematical Society (2001)

[7] Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2), 255–308 (2009)

[8] Carlsson, G., de Silva, V., Morozov, D.: Zigzag persistent homology and real-valued functions. Symposium on Computational Geometry pp. 247–256 (2009)

[9] Carrière, M., Oudot, S.: Local equivalence and intrinsic metrics between Reeb graphs. Symposium on Computational Geometry (2017)

[10] Carrière, M., Oudot, S.: Structure and stability of the one-dimensional mapper. Foundations of Computational Mathematics **18**(6) (2018)

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

[12] Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Extending persistence using Poincaré and Lefschetz duality. Foundations of Computational Mathematics **9**(1), 79–103 (2009)

[13] Dey, T.K., Shi, D., Wang, Y.: Comparing graphs via persistence distortion. Symposium on Computational Geometry **34**, 491–506 (2015)

[14] Edelsbrunner, H., Harer, J.: Persistent homology - a survey. Contemporary Mathematics **453**, 257–282 (2008)

[15] Fabio, B.D., Landi, C.: The edit distance for Reeb graphs of surfaces. Discrete & Computational Geometry **55**(2), 423–461 (2016)

[16] Gasparovic, E., Gommel, M., Purvine, E., Sazdanovic, R., Wang, B., Wang, Y., Ziegelmeier, L.: A complete characterization of the one-dimensional intrinsic Čech persistence diagrams for metric graphs. In: Research in Computational Topology (2018)

[17] Hatcher, A.: Algebraic Topology. Cambridge University Press (2002)

[18] Mémoli, F., Okutan, O.B.: Metric graph approximations of geodesic spaces. arXiv:1809.05566 (2018)

[19] Morozov, D., Beketayev, K., Weber, G.H.: Interleaving distance between merge trees. Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications (2013)

[20] Munkres, J.: Elements of Algebraic Topology. Advanced book classics. Perseus Books (1984)

[21] Oudot, S., Solomon, E.: Barcode embeddings for metric graphs. arXiv:171203630 (2018)

[22] Pirashvili, M., Steinberg, L., Belchi Guillamon, F., Niranjan, M., Frey, J.G., Brodzki, J.: Improved understanding of aqueous solubility modeling through topological data analysis. Journal of Cheminformatics **10**(54) (2018)

[23] de Silva, V., Munch, E., Patel, A.: Categorified Reeb graphs. Discrete & Computational Geometry **55**(4), 854–906 (2016)

[24] Singh, G., Mémoli, F., Carlsson, G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurographics Symposium on Point-Based Graphics (2007)

[25] Sousbie, T.: The persistent cosmic web and its filamentary structure - I. Theory and implementation. Monthly Notices of the Royal Astronomical Society **414**, 350–383 (2011)

[26] Tupin, F., Maitre, H., Mangin, J.F., Nicolas, J.M., Pechersky, E.: Detection of linear features in SAR images: application to road network extraction. IEEE Transactions on Geoscience and Remote Sensing **36**(2), 434–453 (1998)

[27] Umeyama, S.: An eigendecomposition approach to weighted graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **10**(5), 695–703 (1998)

[28] Zeng, Z., Tung, A.K., Wang, J., Feng, J., Zhou, L.: Comparing stars: On approximating graph edit distance. Proceedings of the VLDB Endowment **2**(1), 25–36 (2009)