Local Persistent Homology Based Distance Between Maps *

Mahmuda Ahmed
Dept. of Computer Science
Univ. of Texas at San Antonio
San Antonio, TX, USA
mahmudaahmed@gmail.com

Brittany Terese Fasy
Dept. of Computer Science
Tulane University
New Orleans, LA, USA
brittany@fasy.us

Carola Wenk
Dept. of Computer Science
Tulane University
New Orleans, LA, USA
cwenk@tulane.edu

ABSTRACT

We define a topology-based distance metric between road networks embedded in the plane. This distance measure is based on local persistent homology, and employs a local distance signature that enables identification and visualization of local differences between the road networks. This paper is motivated by the need to recognize changes in road networks over time and to assess the quality of different map construction algorithms. One particular challenge is evaluating the results when no ground truth is known. However, we demonstrate that we can overcome this hurdle by using a statistical technique known as the bootstrap.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Non-numerical Algorithms and Problems

General Terms

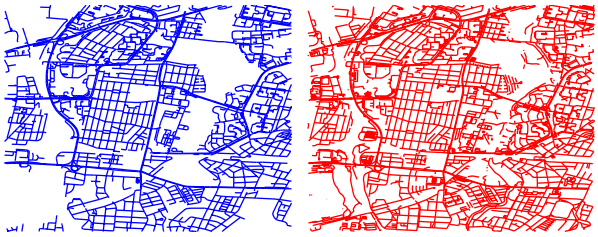
Theory, Algorithms, Experimentation, Measurement

Keywords

Map Comparison, Metrics, Persistence, Local Homology

1. INTRODUCTION

Comparing two graphs embedded in a metric space is important in the field of transportation network analysis. Given street maps of the same city collected from different sources, researchers often need to know how they differ; see Figure 1. The Open Street Map project[1] provides street map and trajectory data open to the public. Street map comparison has received a lot of attention lately with the emergence of algorithms that construct road networks from GPS trajectory data; see e.g. [1, 4, 10, 15, 17, 18]. Only recently, first attempts have been made at comparing results from different map construction algorithms [3, 6]. A few distance measures have been used for this purpose [2, 6, 17], but they are either heuristic in nature or must make strong assumptions to provide theoretical guarantees. Overall, it has remained a challenge to evaluate the quality of the reconstructed networks due to (i) the absence of suitable map comparison algorithms (that may require more than one distance value to capture the similarities and dissimilarities of the maps) and (ii) the lack of applicable ground-truth maps for comparison, especially when the goal of reconstruction is a subgraph of the true road network.



(a) 2007 Road Network (b) 2013 Road Network

Figure 1: These two maps of Berlin were obtained from two different sources, and are dated six years apart. The 2007 map, obtained from TeleAtlas, appears to have less detail than the 2011 map, obtained from OpenStreetMap. One goal of map comparison is to locate where the road network has changed.

Let $G_0 = (V_0, E_0)$ be a road network represented as a graph embedded in a compact metric space $\mathbb{X}$. Although not necessary, we may assume that $\mathbb{X} \subset \mathbb{R}^2$ and that it inherits the Euclidean norm. Suppose $G_1 = (V_1, E_1)$ is another road network embedded in $\mathbb{X}$. We wish to define a distance between G_0 and G_1 that takes a combination of spatial proximity and structural similarity into account. In particular, we make the following contributions:

1. We present a novel topology-based distance measure between road networks.

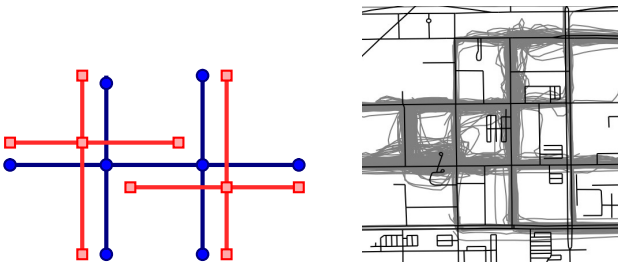
2. We define a local distance signature $\psi_r \colon \mathbb{X} \to \mathbb{R}$, which assigns a non-negative real value to each point $x \in \mathbb{X}$ that captures the difference between G_0 and G_1 as observed by x.

*This work has been partially supported by the National Science Foundation grant CCF-1301911.

[1]www.openstreetmap.org

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666390.



(a) Different Topologies (b) GPS Trajectories

Figure 2: Left: Road networks can have a small Hausdorff distance, but different topological structures. Right: GPS trajectories from university buses in Chicago.

3. We show how to overcome the lack of a known ground-truth road network when evaluating map construction algorithms, by applying a statistical bootstrap to compute a confidence set for the unknown road network.

We prove theoretical properties of our distance measure, and we provide experimental results demonstrating the strengths of this distance. The local distance signature allows us to identify and visualize the local differences between the two graphs. This paper provides the first theoretical foundation for comparing the local topology of two road networks and for defining distance measures for road networks in the absence of an applicable ground-truth road network.

2. MOTIVATION AND RELATED WORK

Given two embedded graphs $G_0 = (V_0, E_0)$ and $G_1 = (V_1, E_1)$, we are interested in computing the distance between them. For example, G_0 could represent the true road network, and G_1 could be the approximation of the true network, computed from GPS trajectory data.

In this section, we outline two distance measures that take the connectivity of the road networks into consideration. First, however, we motivate why we are interested in measuring the distance between road networks.

2.1 Motivation

Graph comparison lies at the core of many applied and theoretical research avenues; see [12] for a review. The two main reasons we are interested in computing a distance between road networks is to rank reconstruction algorithms and to locate changes in road networks.

Ranking Reconstructions. A GPS trajectory T is a sequence of points in $\mathbb{R}^2$ assumed to be close to a road network G; see Figure 2b. Let $T_1, \ldots, T_n$ be n trajectories, which we assume have been sampled i.i.d. from some distribution over the space of all GPS trajectories. We wish to construct $\widehat{G}_n = \widehat{G}_n(T_1, \ldots, T_n)$ as an estimate of G. As mentioned in the introduction, several algorithms exist to reconstruct road networks from this trajectory data; see, e.g., [1, 4, 10, 15, 17, 18]. One goal of having a well-defined distance between graphs is to rank these algorithms.

For a particular known network G_0 and reconstructed estimates of that network $\widehat{G}_n^1, \widehat{G}_n^2, \ldots \widehat{G}_n^k$, we compute the distance between G_0 and $\widehat{G}_n^i$ for $i = 1, 2, \ldots, k$ in order to create a total ordering of the reconstruction algorithms for G_0. Doing this for many different graphs, we can analyze the strengths and weaknesses of different algorithms. For example, perhaps one algorithm outperforms the rest when there is little or no noise in the trajectories, but a different algorithm works best for noisy trajectories.

Detecting Changes in Road Networks. As shown in Figure 1, road networks change over time. One application of this distance metric, and, in particular, of the local distance signature ψ_r is to detect and locate changes in road networks. For example, if roads that previously existed are no longer observed via GPS trajectories, we may want to detect this change. After recognizing this change, we may then investigate why this change occurred.

2.2 Other Distance Measures

One way to measure distance between two embedded graphs is to use the Hausdorff distance; however, this distance does not take the connectivity of a graph into account. On the other hand, *sub-graph isomorphism* and *graph edit distance* calculate the amount of change needed to convert one graph exactly into the second graph. For large graphs, this can be rather expensive. Rather than look at the whole graph, the two distance measures that we describe in this section partially match points (or paths) in one graph to points (or paths) in the other graph. Other algorithms exist, e.g, comparing shortest paths as presented in [17]; however, we only include the algorithms most closely related to the distance metric we propose in this paper in order to highlight the differences between the algorithms.

Hänsel and Gretel Distance. As we mentioned above, the Hausdorff distance does not take the local topology into account. In [6], a sampling-based distance has been proposed that incorporates the local connectivity of the graphs.

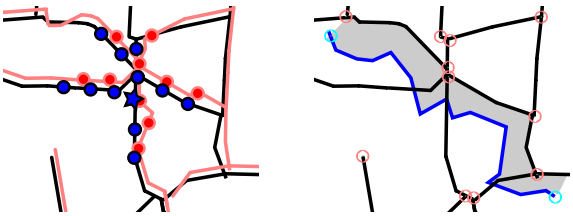
Fix parameters $r > 0$ (locality radius), $d > 0$ (jump distance), and $\delta > 0$ (neighborhood threshold). Then choose a random seed point s in G_0. Next, place a red bread crumb at s, as well as at all points in G_0 at distance kd from s for k an integer and $kd < r$. Here, the distance is measured within G_0. The same process is repeated with graph G_1, placing blue bread crumbs; see Figure 3a. Then a maximum matching between the red bread crumbs and the blue bread crumbs is computed, where a red bread crumb and a blue bread crumb are matched if their distance is at most δ. This results in n_s red bread crumbs and m_s blue ones, of which k_s are matched. Repeating this process for a large number of seeds taken i.i.d., let $n = \sum_s n_s$, $m = \sum_s m_s$, and $k = \sum_s k_s$. Then the *precision* is computed as $\texttt{pre}_{0,1} = k/m$, the *recall* is computed as $\texttt{rec}_{0,1} = k/n$ and the distance is defined using the F-score:

DEFINITION 2.1 (HG-DISTANCE). *The* Hänsel and Gretel (HG) distance *is the statistical F-score, given by*

$$F(G_0, G_1) = 2\frac{\texttt{pre}_{0,1}\texttt{rec}_{0,1}}{\texttt{pre}_{0,1} + \texttt{rec}_{0,1}}.$$

We note here that $F(\cdot, \cdot)$ can take any value in the interval $[0, 1]$. In this measure, higher values indicate a close match between the graph and lower values indicate large differences between the graphs.

This algorithm relies on graph sampling, and is therefore nondeterministic. Specifically, since no convergence guarantees are given, we cannot rely on this distance as a means



(a) Bread Crumbs (b) Map Matching

Figure 3: Left: To compute the HG distance, we must first find a maximal matching between the red and blue breadcrumbs. Right: To compute the PB distance, the blue curve is map-matched onto the black graph.

of ranking reconstruction algorithms. The distance measure that we propose, however, does provide a deterministic distance metric.

Path-Based Distance. In [2], an approach to compute distances between maps has been presented that is based on quantifying how similar or different it is to travel within a road network. Consider the set of paths between two vertices u and v in G_i. A *path* between u and v is the image of a continuous map $\alpha\colon [0,1] \to G_i$ such that $\alpha(0) = u$ and $\alpha(1) = v$. Denote the set of all paths in G_i by Π_i.

The distance measure is based on the Fréchet distance between paths in Π_0 and Π_1. Let $f, g : [0,1] \to \mathbb{R}^2$ be two planar curves. The *Fréchet distance* δ_F between them is

$$\delta_F(f,g) = \inf_{\alpha} \max_{t\in[0,1]} \|f(t) - g(\alpha(t))\|,$$

where $\alpha : [0,1] \to [0,1]$ ranges over all continuous, surjective, non-decreasing reparameterizations. The Fréchet distance is well-suited for comparing paths because it takes continuity and monotonicity of the curves into account.

Definition 2.2 (Path-Based Distance). *The Path-Based Distance between two graphs G_0 and G_1 is defined as*

$$\overrightarrow{d}_{path}(G_0,G_1) = \max_{p_0\in\Pi_0} \min_{p_1\in\Pi_1} \delta_F(p_0,p_1). \qquad (1)$$

By design, the path-based distance is directed and not symmetric, i.e., $\overrightarrow{d}_{path}(G_0,G_1) \neq \overrightarrow{d}_{path}(G_1,G_0)$. This anti-symmetry is desirable, however. For example, G_1 can be the reconstructed road network from bus route data. In this case, the bus routes correspond to a subgraph of the complete road network G_0.

Although this distance is deterministic, it does require a rather strong assumption on the input graphs in order to provide theoretical guarantees. In particular, the requirement that there can be no three-way intersections would present a problem in most road networks.

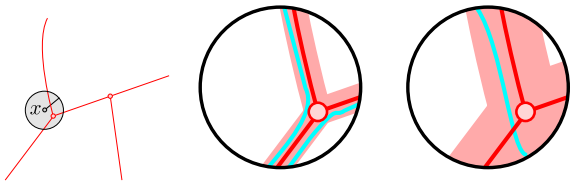
3. LOCAL PERSISTENT HOMOLOGY

We present a distance measure between road networks that compares the local homology of the graphs at different scales. We refer the reader to [5, 14] and Appendix A and the references therein for more details on persistent homology.

In words, the homology (of a graph) describes the connected components as well as the cycles and branching structures present in the graph. The size of these components and cycles, however, is not described by homology alone. Cycles that enclose a small area are given equal importance as cycles that enclose a large area.

In order to assign a size to each cycle in a consistent way, we continuously thicken the graphs until all cycles are filled in. In the process of thickening the graph, new cycles can appear and existing cycles can disappear. The moment that a cycle appears is called the birth of a cycle; the moment a cycle is filled in is called the death of a cycle. In this way, persistent homology defines a set of birth-death pairs, which are plotted in a *persistence diagram*. The scale difference between the birth and the death of a feature is called the lifespan, or the persistence, of a feature. Cycles with long lifespans can be interpreted as important (homological) features of the road network. The distance measure that we use is based on finding a correspondence between the homological features of the graphs.

More formally, let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs embedded in $\mathbb{X} \subset \mathbb{R}^2$. Let $d_{G_1}, d_{G_2}\colon \mathbb{X} \to \mathbb{R}$ be the Euclidean distance functions that map every point in $\mathbb{X}$ to the closest point in G_1 and G_2, respectively. The thickened graphs mentioned above are realized as sublevel sets of these distance functions: $[G_i]^t := d_{G_i}^{-1}((-\infty, t])$.



(a) Neighborhood (b) t_0-thickened (c) t_1-thickened

Figure 4: For each $x \in \mathbb{X}$, we take the intersection of the disc U centered at x (shown in gray) with G (shown in red). We look at the persistence diagram obtained by thickening G and quotienting with $\mathbb{X}-U$. In this example, we see that there are two homology generators at scale t_0 (shown in cyan), but only one at scale t_1.

A sequence of growing topological spaces, such as this sequence of thickened embedded graphs, is a filtration. We look at a special filtration, the local topology filtration, which is defined for a given open neighborhood $U \subset \mathbb{X}$; for example, U can be an open disc centered at a point $x \in \mathbb{X}$. The filtration that we consider is the sequence of topological spaces indexed by t obtained by quotienting the graph thickened by t with the part of the thickened graph not contained in U; see Figure 4. Formally, the local topology filtration is the following continuous sequence of quotient spaces:

$$\mathcal{F}_i(U) = \left\{\mathbb{X}_i^t(U) := [G_i]^t \,/\, ([G_i]^t \cap \mathbb{X} - U)\right\}_{t\geq 0}. \qquad (2)$$

We obtain a set of homology generators with birth and death times (or scales) by computing the persistent homology of this filtration. For simplicity of exposition, we only consider the one-dimensional homology (the relative cycles). We note that we could also consider the zero-dimensional homology (the connected components) with no changes to the theory presented in this paper, and we observe that there are no higher-dimensional generators.

For a fixed scale t, the homology classes of the quotient space $\mathbb{X}_i^t(U)$ correspond to equivalence classes of paths that either begin and end at the same point or begin and end on the boundary of U. In Figure 4(b), the quotient space has two generators (and three non-trivial homology classes). The third homology class can be viewed as the sum of the generating cycles. One of the two generators disappears when the graph is thickened enough to completely cover one of the holes, as shown in Figure 4(c).

Persistent homology, and, in particular, local persistent homology is stable; if $\|d_{G_1} - d_{G_2}\|_\infty$ is small, then the corresponding persistence diagrams will be close. To make this formal, we define a distance between persistence diagrams. The *bottleneck distance* between persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$ is given by $W_\infty(\mathcal{P}_1, \mathcal{P}_2) := \inf_f ||p - f(p)||_\infty$, where $f\colon \mathcal{P}_1 \to \mathcal{P}_2$ ranges over all bijections; see [14, Ch. VII].

Letting $f_1, f_2\colon \mathbb{M} \to \mathbb{R}$ be the functions with persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$, the stability of persistence diagrams can be formally stated as the inequality: $W_\infty(\mathcal{P}_1, \mathcal{P}_2) \leq \|f_1 - f_2\|_\infty$, which was proven by [9, 11] under fairly general conditions. Noticing that (2) is the filtration corresponding to the restriction of d_{G_i} to U relative to ∂U, we have:

Corollary 3.1 (Local Bottleneck Stability). *Let U be an open subset of $\mathbb{X}$. Let $\mathcal{P}_i$ be the persistence diagram corresponding the the filtration $\mathcal{F}_i(U)$, defined in (2). Then, the persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$ are well defined, and $W_\infty(\mathcal{P}_1, \mathcal{P}_2) \leq \|d_{G_1} - d_{G_2}\|_\infty$.*

The analogous results hold for the Wasserstein distance.

4. LOCAL HOMOLOGY BASED DISTANCE

We define a local homology (LH) based distance. We show that it is stable, and provide an approximation algorithm.

4.1 Distances

We compute the persistence diagrams $\mathcal{P}_{i,r}(x)$ corresponding to the filtration $\mathcal{F}_i(B_r(x))$, where $B_r(x)$ is the ball of radius $r > 0$ centered at $x \in \mathbb{X}$. We now use $\mathcal{P}_{1,r}(x)$ and $\mathcal{P}_{2,r}(x)$ to compare the local topologies of G_1 and G_2 near x. We define the *local distance signature* $\psi_r\colon \mathbb{X} \to \mathbb{R}$ by $\psi_r(x) = W_\infty(\mathcal{P}_{1,r}(x), \mathcal{P}_{2,r}(x))$, where $W_\infty(\cdot,\cdot)$ is the Bottleneck distance as defined in the previous section; see Figure 5.

We integrate the local distance signature over $\mathbb{X}$ to obtain a fixed-radius local homology distance:

Definition 4.1 (Fixed Radius Distance). *The fixed-radius local homology distance is:*

$$d_r^{LH}(G_1, G_2) = \int_{\mathbb{X}} \eta(x)\psi_r(x)\, dx,$$

where $\eta\colon \mathbb{X} \to \mathbb{R}$ is a non-negative weight function that integrates to unity.

One weight function of interest is the uniform weight function $\eta(x) = \frac{1}{|\mathbb{X}|}$, where $|\mathbb{X}|$ denotes the Lebesgue measure of $\mathbb{X}$. We note here that d_r^{LH} is not a metric, as observed by the fact that two different functions can have the same persistence diagram. To avoid this issue, we choose a maximum scale $r_1 > 0$ and integrate $d_r^{LH}(\cdot,\cdot)$ over r:

Definition 4.2 (Local Homology Metric). *The local homology distance metric is:*

$$d^{LH}(G_1, G_2) = \int_0^{r_1} \omega(r) \int_{\mathbb{X}} \eta(x)\psi_r(x)\, dx\, dr, \qquad (3)$$

where $\eta\colon \mathbb{X} \to \mathbb{R}$ and $\omega\colon [0, r_1] \to \mathbb{R}$ are non-negative weight functions that integrate to unity.

To compare the local homology based distance against the distance measures mentioned in the previous section, we look at Figures 6 and 7. In Figures 6a and 6b, we show two examples that demonstrate difference that the local homology distance will capture but the Hausdorff distance between graphs does not capture. Figure 7 shows an example where the near-connectivity of a graph is captured by the local homology based distance but not by the path based distance.

Finally, we present one variant of the local homology metric, by integrating over the union of the graphs instead of over the entire domain $\mathbb{X}$. The advantage of using this variant is that we can ignore regions of the domain that are not of interest.

Definition 4.3 (Local Homology Metric Variant). *The Local Homology Metric variant between road networks G_1 and G_2 is the normalized integral of d_r^{LH} over $\mathbb{Y} = G_1 \cup G_2$:*

$$d_{var}^{LH}(G_1, G_2) = \int_0^{r_1} \frac{\omega(r)}{|\mathbb{Y}|} \int_U \psi_r(x)\, dx\, dr, \qquad (4)$$

where $\omega(r)\colon [0, r_1] \to \mathbb{R}$ is a non-negative weight function that integrates to unity and $|\mathbb{Y}|$ denote the total length of edges in $\mathbb{Y}$.

4.2 Properties

In this section, we state the theoretical properties of the LH metric between two embedded graphs; the proofs are deferred to the forthcoming extended version of this paper.

As noted above, d_r^{LH} is not a metric in general; however, it is a pseudometric.

Lemma 4.1 (Existence of Pseudometric). *The fixed-radius distance d_r^{LH} is a pseudometric.*

In fact, the only metric property that d_r^{LH} does not satisfy is the identity of indiscernibles; that is,

$$d_r^{LH}(G_1, G_2) = 0 \Longrightarrow G_1 = G_2$$

does not hold. However, for all $G_1 \neq G_2$, there exists a radius r_0 such that for all $r \leq r_0$, there exists a ball B of radius r that intersects G_1 but not G_2 (or vice versa). Hence, for all $r \leq r_0$, we also have $d_r^{LH}(G_1, G_2) \neq 0$. Integrating over x, we have:

Lemma 4.2 (Metric Space). *If there exists $r_1 > 0$ such that the weight function $\omega(r)$ is positive for all $r < r_1$ and if $\eta(x)$ is positive for all $x \in \mathbb{X}$, then the distance d^{LH} is a metric.*

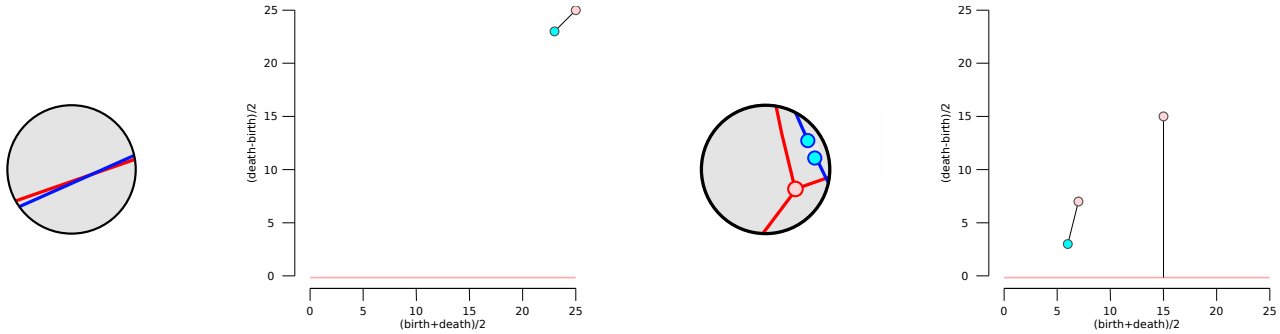
In addition, the variant we present is also a metric:

Lemma 4.3 (Variant is a Metric). *If there exists $r_1 > 0$ such that the weight function $\omega(r)$ is positive for all $r < r_1$, then $d_{var}^{LH}(G_1, G_2)$ is a metric.*

Small changes to graphs G_1 and G_2 should cause small changes in the distance metric. This is, in fact, true:

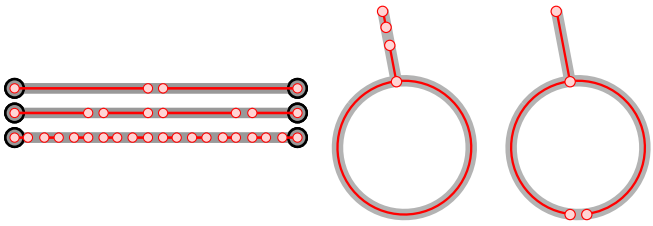
Lemma 4.4 (Stability of Local Homology Metric). *Let d_{G_i} be the distance function to G_i. The local homology distance is stable:*

$$d^{LH}(G_1, G_2) \leq ||d_{G_1} - d_{G_2}||_\infty.$$



(a) Locally Similar (b) Locally Dissimilar

Figure 5: Given two graphs, we obtain two persistence diagrams. We then pair the points in the diagrams, as shown with the solid edges. The local distance signature $\psi_r(x)$ is the length of the longest edge in this pairing. On the left, the local distance signature is small, since the corresponding persistence diagrams are close. On the right, the local distance signature is large.



(a) Multiple Discontinuities (b) Cycles

Figure 6: The Hausdorff distance alone is not strong enough to distinguish between various types of discrepancies between graphs. Left: The Hausdorff distance fails at detecting multiple discrepancies between the pink and gray graphs. Right: The Hausdorff distance fails at detecting if cycles are broken. However, the LH distance could be used to distinguish these five examples.

In order to interpret the local homology distance between two embedded graphs, it is useful to know the set of potential values for d^{LH}.

LEMMA 4.5 (ALLOWABLE DISTANCES). *The following upper bound holds for the local distance signature: $\psi_r(x) \leq \frac{r}{2}$ for all $x \in \mathbb{X}$. Moreover,*

$$0 \leq d^{LH}(G_1, G_2) \leq \frac{r}{2}.$$

4.3 Computation

We approximate the local homology metric in three steps working from the inside of (3) out. For simplicity, we assume that we have uniform weight functions ω and η.

Step One: Computing $\psi_r(x)$. Let $x \in \mathbb{X}$ and $r > 0$. We can compute the homology $H(t)$ of $\mathbb{X}_i^t(U)$ by computing the homology of a corresponding nerve complex; see Appendix A).

We first must compute the persistence diagrams $\mathcal{P}_{i,r}(x)$ for $i = 1, 2$. To do this, let E_i' be a set of circular arcs that cover $\partial B_r(x)$, and such that two arcs intersect in at most one point. Furthermore, we assume that each line containing an edge in E_i intersects each arc of E_i' at most once. We note here that such a subdivision is always possible by dividing $\partial B_r(x)$ into at most $2n + 1$ arcs, where n is the number of edges in G_i. In fact, we can do better than this for small values of r. We consider the set of edges and arcs:

$$\mathcal{K}' = \mathcal{K}_r'(i) = E_i \cup E_i',$$

The homology groups of the nerve $N_t(\mathcal{K}')$ are isomorphic to the homology groups of the quotient space $X_i^t(B_r(x))$; thus, we can use the Nerve filtration to compute the persistence diagrams. In practice, we use the Vietoris-Rips filtration provided in Dionysus.[2]

Step Two: Discretizing the integral over $\mathbb{X}$ for fixed r. Let $\{B_r(x_i)\}_{i=1\ldots N}$ be a finite cover of $\mathbb{X}$. We estimate the integral $d_r^{LH}(G_1, G_2)$ with the following summation:

$$\frac{1}{N}\sum_{i=1}^{N} \psi_r(x_i).$$

Although not necessary, we assume that the points in the set $\{x_i\}$ lie on a lattice. If the points do not lie on a lattice, then we will need to multiply $\psi_r(x_i)$ by a weight that is inversely proportional to the density of the ball centers near x_i.

Step Three: Discretizing the integral over r. In Definition 4.2, the value r_1 is a parameter. It is the largest radius for which we wish to compute the fixed-radius distance. It is likely to be proportional to the maximum distance between corresponding vertices in the input graphs. We assume that r_1 and $k > 0$ are given, then we replace the integral over r by the following summation:

$$\frac{1}{k}\sum_{j=0}^{k}\frac{1}{N}\sum_{i=1}^{N}\widehat{\psi}_{\frac{j}{k}r_1}(x_i).$$

5. THE BOOTSTRAP

One challenge in evaluating map construction algorithms is to evaluate the reconstructed road networks when the

[2]Dionysus is a C++ library for computing persistent homology,developed by Dmitriy Morozov. `http://mrzv.org/software/dionysus/`

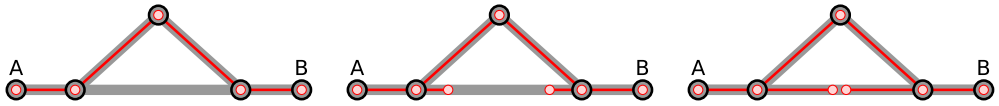


Figure 7: The path-based distance between the gray and the pink graphs is constant for all three examples shown; however, the local-homology distance decreases from left to right.

ground truth is unknown. To overcome this challenge, we employ a statistical technique known as the *bootstrap* [13].

5.1 Correctness of Reconstructions

We would like to say that a road network construction algorithm is correct if the reconstructed road network approaches the true road network as the number of trajectories used to create the reconstruction increases; however, we must be more precise about what this means. Given a graph G, we define the distance function $d_G \colon \mathbb{X} \to \mathbb{R}$ from a point in the domain $\mathbb{X} \subset \mathbb{R}^2$ to the nearest point in the graph G:

$$d_G(x) := \inf_{g \in G} ||x - g||.$$

Given n trajectories, let $\widehat{G}_n$ be a reconstruction of the unknown road network G. We say that $\lim_{n\to\infty} \widehat{G}_n = G$ if $\lim_{n\to\infty} ||d_{\widehat{G}_n} - d_G||_\infty = 0$. Equivalently, we could require that the Hausdorff distance between $\widehat{G}_n$ and G approaches zero: $\lim_{n\to\infty} \mathcal{H}(\widehat{G}_n, G) = 0$. For reasons that will become clear later, we prefer the former definition. If $\widehat{G}_n$ limits to G, then we say that the reconstruction algorithm is correct. We formalize that requirement here, as we will need it later:

Property 5.1 (Correctness of Reconstruction). *If $\widehat{G}_n$ is the road network reconstructed from n i.i.d. trajectories, then:* $\lim_{n\to\infty} \widehat{G}_n = G$.

For the reconstruction algorithm described in [4], the following holds:

Lemma 5.2 (Upper Bound on Approximation Error). *Let $\widehat{G}_n$ be the road network reconstructed using the algorithm in [4]. If $\varepsilon_n/2$ is the maximum map-matching (i.e., Fréchet-matching) distance between G and a trajectory X_i, then:* $\lim_{n\to\infty} ||d_{\widehat{G}_n} - d_G||_\infty \leq \varepsilon_n$.

The term ε_n is the bias of the reconstruction. If, in addition, we let ε_n tend to zero as $n \to \infty$, then we see that [4] satisfies Property 5.1. We approximate the sublevel set filtration of d_G with the sublevel set filtration of $d_{\widehat{G}_n}$.

5.2 Applying the Bootstrap

Let $\mathcal{G}$ be a set of objects, with a distance $d \colon \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ defined. In our case, $\mathcal{G}$ is the set of all road networks embedded in $\mathbb{X} \subset \mathbb{R}$, and d is the fixed-radius local homology distance.

Let G be an unknown road network, with P the probability distribution over the space of all GPS trajectories over G. We sample n points i.i.d. from P and from those n points we construct $\widehat{G}_n \in \mathcal{G}$. We assume $\widehat{G}_n$ is an estimate of an unknown $G \in \mathcal{G}$ and that Property 5.1 is satisfied. Furthermore, we assume that n is large enough that the bias introduced by the estimator $\widehat{G}_n$ is negligible. We are interested in the value of $\delta = d(G, \widehat{G}_n)$, but cannot compute it directly since G is unknown. The distribution P induces a distribution P_δ on distances between G and the random variable $\widehat{G}_n$. For $\alpha \in (0, 1)$, we explain how we use the bootstrap to obtain an upper bound on the random variable δ with probability $1 - \alpha$.

Let $\widehat{P}_n$ denote the uniform distribution over the n sampled trajectories. Draw $X_1^*, \ldots, X_n^*$ i.i.d. from P_n. This is equivalent to sampling n trajectories from the original sample with replacement. Compute $\widehat{G}^*$, the estimated graph computed from $X_1^*, \ldots, X_n^*$. We repeat this process B times (with B very large) to obtain $\widehat{G}_1^*, \ldots, \widehat{G}_B^*$. Now, for each $i \in 1 \ldots B$, we compute $\delta_i = d(\widehat{G}_i^*, \widehat{G}_n)$.

Thinking of δ_i as a random variable, we have the empirical cumulative distribution function $\widehat{F}_n(q) = \frac{1}{B} \sum_{i=1}^{B} I(\delta_i \geq q)$, which approximates the CDF for δ. The upper α quantile of $\widehat{F}$, given by

$$q_\alpha = \inf \left\{ q \ : \ \widehat{F}_n(q) \leq \alpha \right\}$$

can be computed; see Figure 9b. Thus, a $(1 - \alpha)$-confidence interval for the variance of $d(G, \widehat{G}_n)$ is $[0, q_\alpha]$.

6. EXPERIMENTAL RESULTS

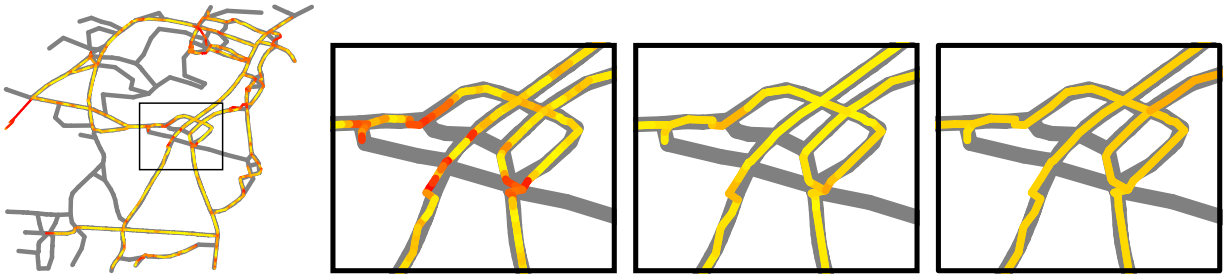
In this section, we give our experimental results. First, we visualize the local distance measure between two reconstructions of the road network in Athens, Greece. Second, we use the bootstrap to compute a confidence set for the distance between a reconstructed road network and a ground truth road network, using GPS trajectories of buses in Chicago, IL. Third, we demonstrate that our distance measure can be used to rank reconstruction algorithms.

Datasets. The *Athens* dataset and source code of reconstruction algorithms from [4], [17] and [2] were downloaded from www.mapconstruction.org. This dataset contains 129 GPS trajectories with a total length of 443.20km (average: 3.82km and standard deviation: 1.45km) obtained from school buses covering an area of 2.6km × 6km; the trajectories range from 13 to 47 position samples, with a sampling rate of 20s to 30s (average: 34.07s and standard deviation: 31.92s). The corresponding ground-truth map, obtained from www.openstreetmap.org, consists of 3436 links (edges) and 2694 nodes. It covers an area of 2.6km × 6km. The edges have a total length of 193km.

The *Chicago* dataset and source code of the algorithm in [7] were downloaded from bits.cs.uic.edu. The *Chicago* dataset contains 889 GPS trajectories with a total length of 2869km (average: 3.22km and standard deviation: 894.28m) obtained from university shuttle buses covering an area of 7km × 4.5km; the trajectories range from 100 to 363 position samples, with a sampling rate of 1s to 29s (average: 3.61s and standard deviation: 3.67s).

6.1 Local Signature

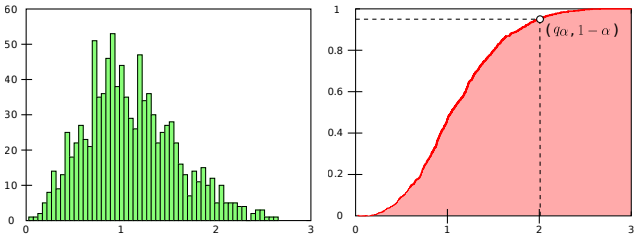
We illustrate the local distance signature with the *Athens* dataset by comparing the road networks obtained from two



(a) Local Homology (b) LH Detailed View (c) Hausdorff (d) Fréchet

Figure 8: We compare two reconstructions of the Athens data set. The thicker gray road network is G_{BE} and the color-coded road network is G_{KP}. The yellow parts of G_{KP} correspond to sections of the graph that have small signatures, and the red parts correspond to large signatures. Notice that the local homology distance is the only one that picks up the locations of the missing intersections.

different reconstruction algorithms. Let G_{BE} be the road network reconstructed using the KDE algorithm found in [7] and G_{KP} be the shortest-path based algorithm found in [17]. Figure 8(a) and (b) illustrate the local distances between the two road networks. In gray is the reconstructed graph G_{KP}. The color-coded road network is the reconstructed graph G_{BE}. The colors represent the local homology signature at $x \in G_{BE}$ for radius $r = 25$m. The local homology signature captures missing intersections, as illustrated in the detailed view of the signatures. We contrast this to the *Hausdorff* and the *Fréchet signatures* shown in Figure 8(c) and (d), respectively. The Hausdorff signature at a point $x \in G_{BE}$ is the distance to G_{KP}. As shown, the Hausdorff signature fails to capture these missing features. A complete comparison between road networks must take topological information into consideration. The Fréchet signature is the distance assigned to a vertex in [2] for link length three paths. To capture the topological differences with confidence, we would need to look at all paths, which is computationally expensive.



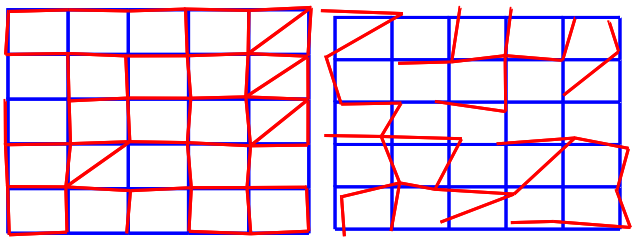
(a) Histogram (b) CDF

Figure 9: We summarize the data from the bootstrap applied to the local homology variant given by (4). On the left is the histogram of the distances from the bootstrap simulation. On the right is the corresponding cumulative distribution function.

6.2 Bootstrap Algorithm

We use the *Chicago* dataset and the iterative road network reconstruction algorithm of [4] in order to demonstrate applying the bootstrap to road network comparison. We resampled the $n = 889$ trajectories $B = 1000$ times with replacement and created graphs $\widehat{G}_1^*, \ldots, \widehat{G}_{1000}^*$ using the iterative algorithm. We use the bootstrap algorithm described in Section 5 in order to estimate the distribution of distances $d_U^{LH}(G, \widehat{G}_n)$ with the distribution of $d_U^{LH}(\widehat{G}_i^*, \widehat{G}_n)$. When computing the bootstrap, we used the local homology variant for a fixed radius $r = 50$ meters. We estimated the integral over the bounding box of the graphs with 1330 balls on a regular grid, giving weight zero to those balls that did not intersect either graph.

The bootstrap works best when the underlying distribution is approximately normally distributed. We look at the estimated distribution in Figure 9A, and confirm that it is approximately normally distributed. The distances obtained in the bootstrap simulation ranged between 0.028084 and 2.655220, with the 95% quantile being 1.983330. Thus, with 95% certainty, the local homology distance between the unknown G and the reconstruction $\widehat{G}_n$ is less than 1.983303.



(a) Small p (b) Large p

Figure 10: The *ground truth* is the blue graph, a regular grid over $[0, 10] \times [0, 10]$. The red graphs represent *reconstructed graphs*, with the quality of the reconstruction decreasing as p increases.

6.3 Comparing Algorithms

In Section 2, we mentioned that the need to rank reconstruction algorithms is one of the motivations for defining a distance measure. In this section, we give experimental evidence that the local homology distance measure outperforms other algorithms at this task. Then, we use the local homology based distance algorithm to rank three road network reconstruction algorithms.

In order to asses the ability of our distance measure to

	PB distance	HG distance	LH distance
Algorithm-1	#3	#3	#1
Algorithm-2	#2	#1	#2
Algorithm-3	#1	#2	#3

Table 1: This table shows the rankings of three road construction algorithms, using three different distance measures: path-based, Hänsel and Gretel, and the local homology based distance. For a full explanation of this table, see Section 6.3.

rank reconstruction algorithms, we created a map G and nine sets of estimations of that map with increasing allowable deviations from G. The map G is a regular grid over $[0, 10] \times [0, 10]$ using the coordinates with even integers as the vertices; see the blue graph in Figure 10.

The perturbation parameter is p, which is allowed to be between zero and one. We perturbed G in three ways to obtain G_p. First, we added diagonals through a cell with probability p. Second, we deleted each edge (including the added diagonal edges) with probability p. Third, we perturbed the vertex at (i, j) by choosing two numbers α and β uniformly at random in the interval $[-p, p]$ and moving the vertex at (i, j) to $(i + \alpha, j + \beta)$. Thus, as p increases, the distance to the ground truth G should increase as well, in expectation. For example, see the red graphs in Figure 10. For each value p, we generate 100 perturbations.

In Figure 11a, we observe that as p increases, so does the median distance observed. Moreover, we see that the middle 50% of the distances do not overlap until $p = 0.8$. This indicates that the LH distance measure can successfully distinguish different values of p. We also compute similar boxplots for two other distance algorithms: the Hänsel and Gretel (HG)-distance and the path-based distance. Recalling that the HG-distance is an F-score, we observe that distance should decrease as p increases. In fact, this is exactly what we see: the median is strictly decreasing. The middle 50% begin to overlap at $p = 0.6$. We interpret this as follows: this controlled experiment illustrates that the HG-distance can be used for ranking, but the local homology based distance is more consistent and more sensitive to perturbations. On the other hand, the path-based distance seems to perform rather poorly at discriminating different values of p.

Having justified the use of the local homology based distance for ranking reconstruction algorithms, we rank three reconstruction algorithms on the Athens dataset. Algorithm-1, given in [17], infers intersection nodes by detecting changes in the direction of movement observed in the GPS trajectories. Edges are inferred by bundling trajectories that are close to two given intersections. Algorithm-2, given in [4], is an incremental algorithm that uses the Fréchet distance between a new trajectory and the closest path in the current road network in order to determine if an update needs to be made. Algorithm-3, given in [6], uses ridges in a kernel density estimate constructed over GPS trajectories in order to determine locations of the roads. See Figure 12 for the illustration of the Athens road network and the three reconstructions, and Table 1 for a summary of the rankings obtained.

Using the local-homology based distance, we find that Algorithm-1 is the closest to the ground truth (with distance 20.65), and that Algorithm-2 the second closest (with distance 20.89 from the ground truth). Algorithm-3 is the farthest from the ground truth, being at distance 21.64.

For comparison, we look at the rankings obtained by using three different distance measures. First, we consider the path-based distance of [2], computing the distance from the reconstruction to the ground truth. Using this distance measure, Algorithm-3 is the closest to the ground truth, being only 73 meters from it. Algorithms-1 and -2 have similar distances of 229 and 224 meters, respectively. The fact that Algorithm-3 is so relatively close to the ground truth can be attributed to the use of kernel density estimates, as well as the fact that Algorithm-3 captures fewer roads from the ground truth.

Using a one-sided Hausdorff distance, we obtain the same rankings as the path-based distance, but with a less severe gap between Algorithm-3 and Algorithms-1 and -2: Algorithm-3 is distance 74 meters from the ground truth, Algorithm-2 is distance 82 meters, and Algorithm-1 is distance 84 meters.

Finally, the Hänsel and Gretel distance of [6] is the only distance measure which ranks Algorithm-1 as the farthest from the ground truth. We compute the HG-distance using $r = 300$ meters, $d = 5$ meters, and $\delta = 100$ meters. (See Section 2.2 for an explanation of these values). The observed F-scores for this graph are 0.42 for Algorithm-2, 0.34 for Algorithm-3, and 0.25 for Algorithm-1. Since a visual inspection would probably rank Algorithm-1 as the best reconstruction, we see that our distance measure is the only measure that has captured the preferred rankings for this data set.

7. CONCLUSION

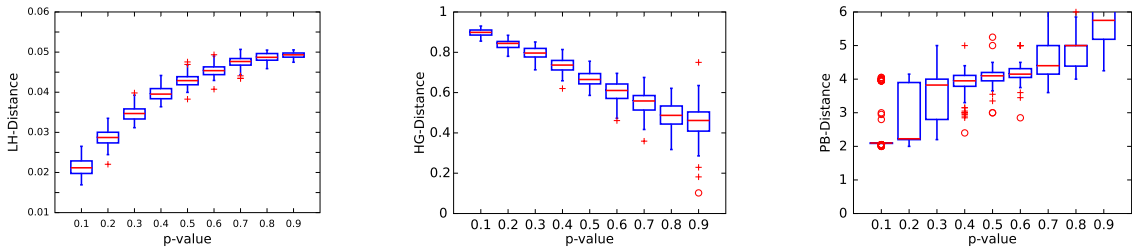
In this paper, we defined a novel distance measure between road networks. Our local homology based distance measure captures both spatial proximity and local topology, by comparing the local persistent homology of the distance functions to the road networks. One unique aspect of the distance measure presented in this paper is that road networks are compared at different *scales*. We have shown how to apply the statistical Bootstrap to give quality guarantees for constructed road networks in the absence of the unknown ground-truth. In addition to providing a single number to quantify the distance between road networks, the local signatures presented in this paper lends itself naturally to the detection and visualization of regions where road networks differ topologically; see Figure 8.

In future work, we will extend the experiment results presented in this paper by analyzing which features are captured by the different distance measures. In addition, we plan to study notions of correctness of road network construction algorithms and the convergence rate for the bootstrap.

Acknowledgements

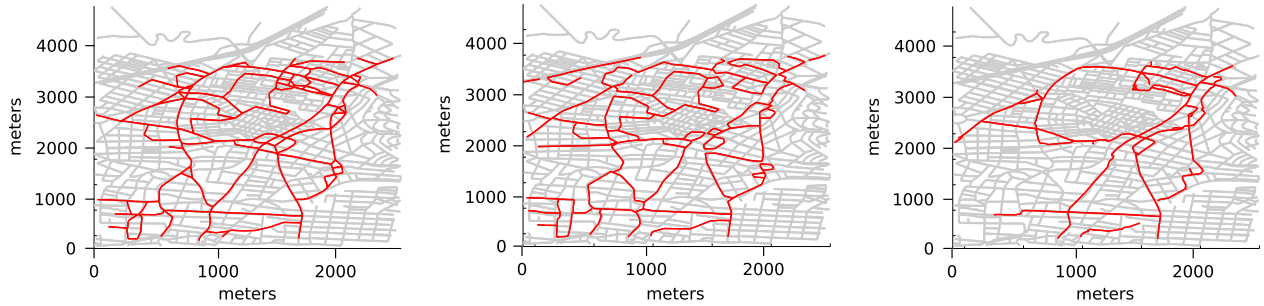
The authors thank James Biagioni for providing his code for map comparison, and Sophia Karagiorgou for help with data conversion and generating maps.

8. REFERENCES



(a) Local Homology Based (b) Hänsel and Gretel (c) Path-Based

Figure 11: For each p-value, we show a boxplot of the distances obtained by generating 100 graphs. For the LH distance (left), we observe that as p increases, the distance increases and the variability decreases. In particular, we notice that the mean distance observed is strictly increasing with p. For the HG distance (middle), we observe that the boxplots seem to decrease (indicating that the perturbations get farther from the ground truth) as p increases. For the path-based distance (right), we observe that consecutive values of p cannot be distinguished very well.



(a) Algorithm-1 (b) Algorithm-2 (c) Algorithm-3

Figure 12: From left to right, the local homology based distance between the reconstruction (shown in gray) and the ground truth (shown in red) is increasing. The x- and y-coordinates are the offsets (in meters) from an arbitrary location, given in Universal Transverse Mercator (UTM) coordinates. That location is UTM Zone 34S, 482,400 meters east, 4,213,250 meters north.

[1] Aanjaneya, M., Chazal, F., Chen, D., Glisse, M., Guibas, L., and Morozov, D. Metric graph reconstruction from noisy data. *Internat. J. Comput. Geom. Appl. 22*, 04 (2012), 305–325.

[2] Ahmed, M., Fasy, B. T., Hickmann, K. S., and Wenk, C. Path-based distance for street map comparison, 2014. arXiv:1309.6131.

[3] Ahmed, M., Karagiorgou, S., Pfoser, D., and Wenk, C. A comparison and evaluation of map construction algorithms, 2014. arXiv:1402.5138.

[4] Ahmed, M., and Wenk, C. Constructing street networks from gps trajectories. In *Algorithms – ESA 2012* (Sept. 2012), L. Epstein and P. Ferragina, Eds., vol. 7501 of *Lect. Notes Comput. Sci.*, Springer Berlin-Heidelberg.

[5] Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Morozov, D. Inferring local homology from sampled stratified spaces. In *IEEE Symp. Found. Comput. Sci.* (2007), IEEE, pp. 536–546.

[6] Biagioni, J., and Eriksson, J. Inferring road maps from global positioning system traces. *J. Transportation Res. Board 2291*, 1 (2012), 61–71.

[7] Biagioni, J., and Eriksson, J. Map inference in the face of noise and disparity. In *Proc. of the Twentieth ACM SIGSPATIAL GIS* (2012), ACM, pp. 79–88.

[8] Borsuk, K. On the imbedding of systems of compacta in simplicial complexes. *Fund. Math. 35*, 1 (1948), 217–234.

[9] Chazal, F., de Silva, V., Glisse, M., and Oudot, S. The structure and stability of persistence modules, July 2012. arXiv:1207.3674.

[10] Chen, D., Guibas, L. J., Hershberger, J., and Sun, J. Road netowrk reconstruction for organizing paths. In *Proc. of the* 21^{st} *ACM-SIAM Symp. Discrete Algorithms* (2010), SIAM, pp. 1309–1320.

[11] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete Comput. Geom. 37*, 1 (2007), 103–120.

[12] Conte, D., Foggia, P., Sansone, C., and Vento, M. Thirty years of graph matching in pattern

recognition. *Internat. J. Artif. Intell. 18*, 3 (2004), 265–298.

[13] Davison, A. C., and Hinkley, D. V. *Bootstrap Methods and Their Application*, vol. 1. Cambridge UP, 1997.

[14] Edelsbrunner, H., and Harer, J. *Computational Topology. An Introduction.* Amer. Math. Soc., Providence, RI, 2010.

[15] Ge, X., Safa, I., Belkin, M., and Wang, Y. Data skeletonization via Reeb graphs. In *Proc. of the Eighteenth Annu. Conf. Neural Info. Proc. Sys.* (2011), pp. 837–845.

[16] Hatcher, A. *Algebraic Topology.* Cambridge Univ. P., 2002. Electronic Version.

[17] Karagiorgou, S., and Pfoser, D. On vehicle tracking data-based road network generation. In *Proc. of the Twentieth ACM SIGSPATIAL GIS* (2012), ACM, pp. 89–98.

[18] Liu, X., Biagioni, J., Eriksson, J., Wang, Y., Forman, G., and Zhu, Y. Mining large-scale, sparse GPS traces for map inference: comparison of approaches. KDD, ACM, pp. 669–677.

[19] Munkres, J. R. *Topology.* Prentice Hall, Upper Saddle River, NJ, 1975.

APPENDIX

This appendix has been abbreviated for space constraints. Please refer to the forthcoming full paper for more details, including full proofs of the theorems stated in this short paper.

A. COMPUTATIONAL TOPOLOGY

We refer the reader to [16, 19] for an introduction to topology, and use this appendix to define the nerve of a set of compact sets.

A.1 Compact Sets and Nerves

A topological space is called compact if every open cover has a finite subcover. In particular, a subspace of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. Let $\mathcal{K}$ be a collection of compact sets. We create a simplicial complex from $\mathcal{K}$:

Definition A.1 (Nerve). *The nerve of $\mathcal{K}$ is the simplicial complex*

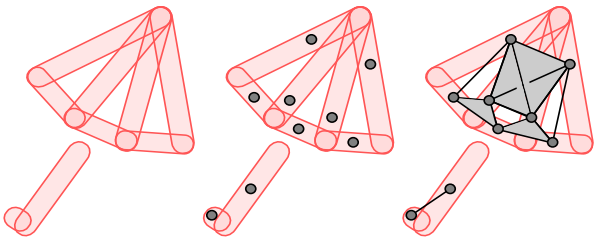
$$N(\mathcal{K}) := \{\mathcal{S} \subseteq \mathcal{K} \ : \ \cap \mathcal{S} \neq 0\},$$

where $\cap\mathcal{S}$ denotes the intersection of all sets in $\mathcal{S}$.

In words, every set in $\mathcal{K}$ corresponds to a vertex in $\mathrm{N}(\mathcal{K})$ and each collection of sets $\mathcal{S}$ corresponds to a simplex if the common intersection of all sets in $\mathcal{S}$ is nonempty; see Figure 13. Under the right assumptions, the nerve of $\mathcal{K}$ captures the topology of $|\cup\mathcal{K}|$; in particular:

Lemma A.1 (Nerve Lemma [8]). *Let $\mathcal{K}$ be a collection of compact sets such that for all $\mathcal{S} \subseteq \mathcal{K}$, we have $\cap\mathcal{S}$ is either empty or homotopic to a point. Then, the topological space $|\cup\mathcal{K}|$ is homotopically equivalent to $N(\mathcal{K})$.*

In particular, the homology groups $H_p(|\mathcal{K}|)$ and $H_p(N(\mathcal{K}))$ are isomorphic.



(a) Compact Sets (b) Vertices (c) Nerve

Figure 13: On the left, we have a collection of sets. In the middle, we add a vertex for each set. On the right, we complete the simplicial complex by adding a simplex for each collection of intersecting sets.

Given a road network $G = (V, E)$, we add a scale parameter $t > 0$ to obtain a set of topological spaces indexed by t:

$$E^t := \{[e]^t \ : \ e \in E\},$$

where $[e]^t$ denotes the set of points at most distance t from the edge e, i.e., the Minkowski sum of e with the ball of radius t. In words, we thicken each edge in E by t. When the sets comprising E lie in $\mathbb{R}^2$, we only need to check subsets up to size three in order to recover the homology of the union of the elements in E^t. Therefore, we describe next how to determine the one-skeleton and the two-skeleton of $N_r(\mathcal{K})$.

One-Skeleton. Let e_1 and e_2 be edges in G. To determine if $\{e_1, e_2\} \in N_t(E)$, we must determine if $([e_1]^t) \cap ([e_2]^t)$ is nonempty. The data structure we use is a matrix D, where the rows and columns are indexed by the edges in G and the ij-entry d_{ij} is the distance between the ith and jth edges. We may think of D as a function $D\colon E \times E \to \mathbb{R}$. The one-skeleton of $N_t(E)$ has a zero-simplex for each edge in E and a one-simplex for each pair e_i, e_j where $i < j$ and $d_{ij} \leq t$.

Two-Skeleton. Let e_1, e_2, and e_3 be edges in G. To determine if $\{e_1, e_2, e_3\} \in N_t(E)$, we must determine if $[e_1]^t \cap [e_2]^t \cap [e_3]^t$ is nonempty. We consider now the three-dimensional matrix $T\colon E \times E \times E \to \mathbb{R}$ whose ijk-entry t_{ijk} is the smallest t for which the intersection is nonempty.

Computing t_{ijk} is equivalent to computing the distance from an edge to the Voronoi vertex in the arrangement of the edges e_1, e_2, and e_3. To find the Voronoi vertex, we intersect two bisectors. Generically, the bisector between two edges e_i and e_j is a continuous curve defined by two parabolas and a line segment. Given two bisectors b_{ij} and b_{jk}, we compute the finite intersection set $b_{ijk} := b_{ij} \cap b_{jk}$. By definition, each b_{ijk} is the set of all points equidistant to all three edges; we record the minimum such distance as t_{ijk}. The triangle corresponding to the set $\{e_1, e_2, e_3\}$ is in $N_t(E)$ if $t_{ijk} \leq t$.

The matrices D and T are sufficient to construct the two-skeleton of $N_t(E)$. In particular, the two skeleton of $N_t(E)$ is equal to the one-skeleton of $N_t(E)$ plus a triangle for every triple of edges e_i, e_j, e_k for which $t_{ijk} \leq t$.