Gromov-Hausdorff Approximation of Filament Structure Using Reeb-type Graph

Frédéric Chazal* and Ruqi Huang † and Jian Sun‡

July 26, 2014

Abstract

In many real-world applications data appear to be sampled around 1-dimensional filamentary structures that can be seen as topological metric graphs. In this paper we address the metric reconstruction problem of such filamentary structures from data sampled around them. We prove that they can be approximated, with respect to the Gromov-Hausdorff distance by well-chosen Reeb graphs (and some of their variants) and provide an efficient and easy to implement algorithm to compute such approximations in almost linear time. We illustrate the performances of our algorithm on a few data sets.

1 Introduction

Motivation. With the advance of sensor technology, computing power and Internet, massive amounts of geometric data are being generated and collected in various areas of science, engineering and business. As they are becoming widely available, there is a real need to analyze and visualize these large scale geometric data to extract useful information out of them. In many cases this data is not embedded in Euclidean spaces and come as (finite) sets of points with pairwise distances information, i.e. (discrete) metric spaces. A large amount of research has been done on dimensionality reduction, manifold learning and geometric inference for data embedded in, possibly high dimensional, Euclidean spaces and assumed to be concentrated around low dimensional manifolds [**?**, **?**, **?**]. However, the assumption of data lying on a manifold may fail in many applications. In addition, the strategy of representing data by points in Euclidean space may introduce large metric distortions as the data may lie in highly curved spaces, instead of in flat Euclidean space raising many difficulties in the analysis of metric data. In the past decade, with the development of topological methods in data analysis, new theories such as topological persistence (see, for example, [**?**, **?**, **?**, **?**]) and new tools such as the Mapper algorithm [**?**] have given rise to new algorithms to extract and visualize geometric and topological information from metric data without the need of an embedding into an Euclidean space. In this paper we focus on a simple but important setting where the underlying geometric structure approximating the data can be seen as a branching filamentary structure i.e., more precisely, as a *metric graph* which is a topological graph endowed with a length assigned to each edge. Such structures appear naturally in various real-world data such as collections of GPS traces collected by vehicles on a road network, earthquakes distributions that concentrate around geological faults, distributions of galaxies in the universe, networks of blood vessels in anatomy or hydrographic networks in geography just to name a few. It is thus appealing to try to capture such filamentary structures and to approximate the data by metric graphs that will summarize the metric and allow convenient visualization. **Contribution.** In this paper we address the metric reconstruction problem for filamentary structures. The input of our method and algorithm is a metric space (X, d_X) that is

*INRIA Saclay - France - frederic.chazal@inria.fr
†INRIA Saclay - France - ruqi.huang@inria.fr
‡Tsinghua University - China - jsun@math.tsinghua.edu.cn

assumed to be close with respect to the so-called Gromov-Hausdorff distance d_{GH} to a much simpler, but unknown, metric graph $(G', d_{G'})$. Our algorithm outputs a metric graph (G, d_G) that is proven to be close to (X, d_X) in both geometry and topology. Our approach relies on the notion of Reeb graph (and some variants of it introduced in Section **??**) and our main theoretical results are stated in the following two theorems.

Theorem 4.9 [Recovery of Geometry]. *Let (X, d_X) be a compact connected geodesic space, let $r \in X$ be a fixed base point such that the metric Reeb graph (G, d_G) of the function $d = d_X(r, .) : X \to \mathbb{R}$ is a finite graph. If for a given $\varepsilon > 0$ there exists a finite metric graph $(G', d_{G'})$ such that $d_{GH}(X, G') < \varepsilon$ then we have*

$$d_{GH}(X, G) < 2(\beta_1(G) + 1)(17 + 8N_{E,G'}(8\varepsilon))\varepsilon$$

where $N_{E,G'}(8\varepsilon)$ is the number of edges of G' of length at most 8ε and $\beta_1(G)$ is the first Betti number of G, i.e. the number of edges to remove from G to get a spanning tree. In particular if X is at distance less than ε from a metric graph with shortest edge larger than 8ε then $d_{GH}(X, G) < 34(\beta_1(G) + 1)\varepsilon$.

Note that $\beta_1(G) \leqslant \beta_1(X)$ and thus $d_{GH}(X, G)$ is upper bounded by the quantities depending only on the input X.

Theorem 5.1 [Recovery of Topology]. *Let (X, d_X) be a compact connected path metric space and $(G', d_{G'})$ is a metric graph so that $d_{GH}(X, G') < \varepsilon$. Let $r \in X$, $\alpha > 60\varepsilon$ and $\mathcal{I}\{[0, 2\alpha), (i\alpha, (i + 2)\alpha) | 1 \leqslant i \leqslant m\}$ covers the segment $[0, Diam(X)]$ such that the α-Reeb graph G associated to $\mathcal{I}$ and the function $d = d_X(r, .) : X \to \mathbb{R}$ is a finite graph. If no edges of G' is shorter than L and no loops of G' is shorter than $2L$ with $L \geqslant (24\alpha + 9\varepsilon)$, then we have G and G' are homotopy equivalent.*

To turn this result into a practical algorithm we address two issues:

(1) Raw data usually do not come as geodesic spaces. They are given as discrete sets of point (and thus not connected metric spaces) sampled from the underlying space (X, d_X). Moreover in many cases only distances between nearby points are known. A geodesic space (see Section 2 for a definition of geodesic space) can then be obtained from these raw data as a neighborhood graph where nearby points are connected by edges whose length is equal to their pairwise distance. The shortest path distance in this graph is then used as the metric. In our experiments we use this new metric as the input of our algorithm. The question of the approximation of the metric on X by the metric induced on the neighborhood graphs is out of the scope of this paper.

(2) Approximating the Reeb graph (G, d_G) from a neighborhood graph is usually not obvious. If we compute the Reeb graph of the distance function to a given point defined on the neighborhood graph we obtain the neighborhood graph itself and do not achieve our goal of representing the input data by a simple graph. See Table 1. It is then appealing to build a two dimensional complex having the neighborhood graph as 1-dimensional skeleton and use the algorithm of [**?**, **?**] to compute the Reeb graph of the distance to the root point. Unfortunately adding triangles to the neighborhood graph may widely change the metric between the data points on the resulting complex and significantly increase the complexity of the algorithm. We overcome this issue by introducing a variant of the Reeb graph, the α-Reeb graph, inspired from [**?**] and related to the recently introduced notion of graph induced complex [**?**], that is easier to compute than the Reeb graph but also comes with approximation guarantees (see Theorem 4.10). As a consequence our algorithm runs in almost linear time (see Section 6).

Related work. Approximation of data by 1-dimensional geometric structures has been considered by different communities. In statistics, several approaches have been proposed to address the problem of detection and extraction of filamentary structures in point cloud data. For example Arial-Castro et al [**?**] use multiscale anisotropic strips to detect linear structure while [**?**, **?**] and more recently [**?**] base their approach upon density gradient descents or medial axis techniques. These methods apply to data corrupted by outliers embedded in Euclidean spaces and focus on the inference of individual filaments without focus on the global geometric structure of the filaments network.

In computational geometry, the curve reconstruction problem from points sampled on a curve in an euclidean space has been extensively studied and several efficient algorithms have been proposed [**?**, **?**,

?]. Unfortunately, these methods restricts to the case of simple embedded curves (without singularities or self-intersections) and hardly extend to the case of topological graphs. In a more intrinsic setting where data come as finite abstract metric spaces, [?] propose an algorithm that outputs a topologically correct (up to a homeomorphism) reconstruction of the approximated graph. However this algorithm requires some tedious parameters tuning and relies on quite restrictive sampling assumptions. When these conditions are not satisfied, the algorithm may fail and not even outputs a graph. Compared to the algorithm of [?], our algorithm not only comes with metric guarantees but also whatever the input data is, it always outputs a metric graph and does not require the user to choose any parameters. Closely related to our approach is the data skeletonization algorithm proposed in [?] that computes the Reeb graph of an approximation of the distance function to a root point on a 2-dimensional complex built on top of the data whose size might be significantly larger than a neighboring graph. The algorithm of [?] also always output a graph but it does not come with metric guaranties. Recently, Bauer, Ge and Wang [?] define a metric based on the function for Reeb graph and show it is stable under Gromov-Hausdorff distance. The implementation of our algorithm relies on the Mapper algorithm [?], that provides a way to visualize data sets endowed with a real valued function as a graph, where the considered function is the distance to the chosen root point. However, unlike the general Mapper algorithm, our methods provides an upper bound on the Gromov-Hausdorff distance between the reconstructed graph and the underlying space from which the data points have been sampled.

In theoretical computer science, there is much of work on approximating metric spaces using trees [?, ?, ?] or distribution of trees [?, ?] where the trees are often constructed as spanning trees possibly with Steiner points. Our approach is different as our reconstructed graph or tree is a quotient space of the original metric space where the metric only gets contracted (see Proposition 4.5). Finally we remark that the recovery of filament structure is also studied in various applied settings, including road networks [?, ?], galaxies distributions [?].

The paper is organized as follows. The basic notions and definitions used throughout the paper are recalled in Section 2. The Reeb and α-Reeb graphs endowed with a natural metric are introduced in Section ?? and the approximation results are stated and proven in Sections 4.1 and 4.2. Our algorithm is described in Section 6 and experimental results are presented and discussed in Section 7.

2 Preliminaries

Recall that a metric space is a pair (X, d_X) where X is a set and $d_X : X \times X \to \mathbb{R}$ is a non negative map such that for any $x, y, z \in X$, $d_X(x, y) = 0$ if and only if $x = y$, $d_X(x, y) = d_X(y, x)$ and $d_X(x, z) \leqslant d_X(x, y) + d_X(y, z)$. Two compact spaces (X, d_X) and (Y, d_Y) are isometric if there exits a bijection $\varphi : X \to Y$ that preserves the distances, namely: for any $x, x' \in X$, $d_Y(\varphi(x), \varphi(x')) = d_X(x, x')$. The set of isometry classes of compact metric spaces can be endowed with the Gromov-Hausdorff distance that can be defined using the following notion of correspondence ([?] Def. 7.3.17).

Definition 2.1. Let (X, d_X) and (Y, d_Y) be two compact metric spaces. Given $\varepsilon > 0$, an *ε-correspondence* between (X, d_X) and (Y, d_Y) is a subset $C \subset X \times Y$ such that: i) for any $x \in X$ there exists $y \in Y$ such that $(x, y) \in C$; ii) for any $y \in Y$ there exists $x \in X$ such that $(x, y) \in C$; iii) for any $(x, y), (x', y') \in C$, $|d_X(x, x') - d_Y(y, y')| \leqslant \varepsilon$.

Definition 2.2. The *Gromov-Hausdorff distance* between two compact metric spaces (X, d_X) and (Y, d_Y) is defined by

$$d_{GH}(X, Y) = \frac{1}{2} \inf\{\varepsilon \geqslant 0 : \text{there exists an } \varepsilon\text{-correspondence between } X \text{ and } Y\}$$

A metric space (X, d_X) is a *path metric space* if the distance between any pair of points is equal to the infimum of the lengths of the continuous curves joining them [1]. In the sequel of the paper we

[1] see [?] Chap.1 for the definition of the length of a continuous curve in a general metric space

consider compact path metric spaces. It follows from the Hopf-Rinow theorem (see [**?**] p.9) that such spaces are *geodesic*, i.e. for any pair of point $x, x' \in X$ there exists a minimizing geodesic joining them.[2] A continuous path $\delta : I \to X$ where I is a real interval or the unit circle is said to be *simple* if it is not self intersecting, i.e. if δ is an injective map.

Recall that a *(finite) topological graph* $G = (V, E)$ is the geometric realization of a (finite) 1-dimensional simplicial complex with vertex set V and edge set E. If moreover each 1-simplex $e \in E$ is a metric edge, i.e. $e = [a, b] \subset \mathbb{R}$, then the graph G inherits from a metric d_G which is the unique one whose restriction to any $e = [a, b] \in E$ coincides with the standard metric on the real segment $[a, b]$. Then (G, d_G) is a *metric graph* (see [**?**], Section 3.2.2 for a more formal definition). Intuitively, a metric graph can be seen as a topological graph with a length assigned to each of its edges.

The *first Betti number* $\beta_1(G)$ of a finite topological graph G is the rank of the first homology group of G, or equivalently, the number of edges to remove from G to get a spanning tree.

3 Reeb-type Graph

In this section, we describe a construction to build a Reeb-type graph for approximating the metric space. Let (X, d_X) be a compact geodesic space and let $r \in X$ be a fixed base point. Let $d : X \to \mathbb{R}$ be the distance function to r, i.e., $d(x) = d_X(r, x)$.

The Reeb graph. The relation $x \sim y$ if and only if $d(x) = d(y)$ and x, y are in the same path connected component of $d^{-1}(d(x))$ is an equivalence relation. The quotient space $G = X/\sim$ is called the *Reeb graph* of d and we denote by $\pi : X \to G$ the quotient map. Notice that π is continuous and as X is path connected, G is path connected. The function d induces a function $d_* : G \to \mathbb{R}_+$ that satisfies $d = d_* \circ \pi$. The relation defined by: for any $g, g' \in G$, $g \leqslant_G g'$ if and only if $d_*(g) \leqslant d_*(g')$ and there exist a continuous path γ in G connecting g to g' such that $d \circ \gamma$ is non decreasing, makes G a partially ordered set.

The α-Reeb graphs. Computing or approximating the Reeb graph of (X, d) from a finite set of point sampled on X is usually a difficult task. To overcome this issue we also consider a variant of the Reeb graph that shares very similar properties than the Reeb graph. Let $\alpha > 0$ and let $\mathcal{I} = \{I_i\}_i \in I$ be a covering of the range of d by open intervals of length at most α. The transitive closure of the relation $x \sim_\alpha y$ if and only if $d(x) = d(y)$ and x, y are in the same path connected component of $d^{-1}(I_i)$ for some interval $I_i \in \mathcal{I}$ is an equivalence relation that is also denoted by $\sim_\alpha$. The quotient space $G_\alpha = X/\sim_\alpha$ is called the *α-Reeb graph*[3] of d and we denote by $\pi : X \to G_\alpha$ the quotient map. Notice that π is continuous and as X is path connected, G_α is path connected. The function d induces a function $d_* : G_\alpha \to \mathbb{R}_+$ that satisfies $d = d_* \circ \pi$. The relation defined by: for any $g, g' \in G_\alpha$, $g \leqslant_{G_\alpha} g'$ if and only if $d_*(g) \leqslant d_*(g')$ and there exist a continuous path γ in G_α connecting g to g' such that $d \circ \gamma$ is non decreasing, makes G_α a partially ordered set.

The α-Reeb graph is closely related to the graph constructed by the Mapper algorithm introduced in [**?**] making its computation much more easier than the Reeb graph (see Section 6).

Notice that without making assumptions on X and d, in general G and G_α are not finite graphs. However when the number of path connected components of the level sets of d is finite and changes only a finite number of times then the Reeb graph turns out to be a finite directed acyclic graph. Similarly, when the covering of X by the connected components of $d^{-1}(I_i), i \in \mathcal{I}$ is finite, the α-Reeb graph also turns out to be a finite directed acyclic graph. This happens in most applications and for example when (X, d_X) is a finite simplicial complex or a compact semialgebraic (or more generally a compact subanalytic space) with d being semi-algebraic (or subanalytic).

[2] recall that a minimizing geodesic in X is any curve $\gamma : I \to X$, where I is a real interval, such that $d_X(\gamma(t), \gamma(t')) = |t - t'|$ for any $t, t' \in I$.

[3] strictly speaking we should call it the α-Reeb graph associated to the covering $\mathcal{I}$ but we assume in the sequel that some covering $\mathcal{I}$ has been chosen and we omit it in notations

All the results and proofs presented in Section 4 are exactly the same for the Reeb and the α-Reeb graphs. In the following paragraph and in Section 4.1, G denotes indifferently the Reeb graph or an α-Reeb graph for some $\alpha > 0$. We also always assume that X and d (and α and $\mathcal{I}$) are such that G is a finite graph.

A metric on Reeb and α-Reeb graphs. Let define the set of vertices V of G as the union of the set of points of degree not equal to 2 with the set of local maxima of d_* over G, and the base point $\pi(r)$. The set of edges E of G is then the set of the connected components of the complement of V. Notice that $\pi(r)$ is the only local (and global) minimum of d_*: since X is path connected, for any $x \in X$ there exists a geodesic γ joining r to x along which d is increasing; d_* is thus also increasing along the continuous curve $\pi(\gamma)$, so $\pi(x)$ cannot be a local minimum of d_*. As a consequence d_* is monotonic along the edges of G. We can thus assign an orientation to each edge: if $e = [p, q] \in G$ is such that $d_*(p) < d_*(q)$ then the positive orientation of e is the one pointing from p to q. Finally, we assign a metric to G. Each edge $e \in E$ is homeomorphic to an interval to which we assign a length equal to the absolute difference of the function d_* at two endpoints. The distance between two points p, p' of e is then $|d_*(p) - d_*(p')|$. This makes G a metric graph (G, d_G) isometric to the quotient space of the union of the intervals isometric to the edges by identifying the endpoints if they correspond to the same vertex in G. Note that d_* is continuous in (G, d_G) and for any $p \in G$, $d_*(p) = d_G(\pi(r), p)$. Indeed this is a consequence of the following lemma.

Lemma 3.1. *If δ is a path joining two points $p, p' \in G$ such that $d_* \circ \delta$ is strictly increasing then δ is a shortest path between p and p' and $d_G(p, p') = d_*(p') - d_*(p)$.*

Proof. As $d_* \circ \delta$ is strictly increasing, when δ enters an edge e by one of its end points, either it exits at the other end point or it stops at p' if $p' \in e$. Moreover δ cannot go through a given edge more than one time. As a consequence δ can be decomposed in a finite sequence of pieces $e_0 = [p, p_1], e_1 = [p_1, p_2], \cdots, e_{n-1} = [p_{n-1}, p_n], e_n = [p_n, p']$ where e_0 and e_n are the segments joining p and p' to one of the endpoint of the edges that contain them and $e_1, \cdots, e_{n-1}$ are edges. So, the length of δ is equal to $(d_*(p_1) - d_*(p)) + (d_*(p_2) - d_*(p_1)) + \cdots + (d_*(p') - d_*(p_n)) = d_*(p') - d_*(p)$ and $d_G(p, p') \leqslant d_*(p') - d_*(p)$.

Similarly any simple path joining p to p' can be decomposed in a finite sequence of pieces $e'_0 = [p, p'_1], e'_1 = [p'_1, p'_2], \cdots, e'_{k-1} = [p'_{k-1}, p'_k], e'_k = [p'_k, p']$ where e'_0 and e'_k are the segments joining p and p' to one of the endpoint of the edges that contain them, and $e'_1, \cdots, e'_{k-1}$ are edges. Now, as we do not know that d_* is increasing along this path, its length is thus equal to $|d_*(p'_1) - d_*(p)| + |d_*(p'_2) - d_*(p'_1)| + \cdots + |d_*(p') - d_*(p'_n)| \geqslant d_*(p') - d_*(p)$. So, $d_G(p, p') \geqslant d_*(p') - d_*(p)$. $\square$

4 Approximation of Metric

4.1 Bounding the Gromov-Hausdorff distance between X and G

The goal of this section is to provide an upper bound of the Gromov-Hausdorff distance between X and G that only depends on the first Betti number $\beta_1(G)$ of G and the maximal diameter M of the level sets of π. An upper bound of M is given in the next section.

Theorem 4.1. *$d_{GH}(X, G) < (\beta_1(G) + 1)M$ where $d_{GH}(X, G)$ is the Gromov-Hausdorff distance between X and G, $\beta_1(G)$ is the first Betti number of G and $M = \sup_{p \in G}\{diam(\pi^{-1}(p))\}$ is the supremum of the diameters of the level sets of π.*

Remark that as $\beta_1(G) \leqslant \beta_1(X)$, from the above theorem, $d_{GH}(X, G)$ is upper bounded by the quantities depending only on the input X. The proof of Theorem 4.1 is deduced from two propositions comparing the distances between pairs of points $x, y \in X$ and their images $\pi(x), \pi(y) \in G$ whose proofs rely on the notion of merging vertex. A vertex $v \in V$ is called a *merging vertex* if it is the end point of at least two edges e_1 and e_2 that are pointing to it according to the orientation defined in Section **??**. Geometrically this means that there are at least two distinct connected components of $\pi^{-1}(d_*^{-1}(d_*(v) - \varepsilon))$ that accumulate to $\pi^{-1}(v)$ as $\varepsilon > 0$ goes to 0. The set of merging vertices is denoted by V_m. We have

Lemma 4.2. *The cardinality of V_m is at most $\beta_1(G)$ where $\beta_1(G)$ is the rank of the first homology group of G.*

Proof. The result follows from classical homology persistence theory [**?**]. First remark that, as $\pi(r)$ is the only local minimum of d_*, the sublevel sets of the function $d_* : G \to \mathbb{R}_+$ are all path connected. Indeed if $\pi(x), \pi(y) \in G$ are in the same sublevel set $d_*^{-1}([0, \alpha])$, $\alpha > 0$, then the images by π of the shortest paths in X connecting x to r and y to r are contained in $d_*^{-1}([0, \alpha])$ and their union is a continuous path joining $\pi(x)$ to $\pi(y)$. As a consequence, the 0-dimensional persistence of d_* is trivial. So as we increase the α value, no merging vertices serve as connecting two different connected components. Thus, each merging vertex in V_m creates at least a cycle that never dies as G is one dimensional and does not contain any 2-dimensional simplex. Thus $|V_m| \leqslant \beta_1(G)$. □

The following lemma show that a shortest path in G is the projection of a shortest path in X as long as it does not meet a merging vertex and allow to prove proposition 4.4 below.

Lemma 4.3. *Let $p, p' \in G$ and let $\delta : [d_*(p), d_*(p')] \to G$ be a strictly increasing path going from p to p' that does not contain any point of V_m in its interior. Then for any $x' \in \pi^{-1}(p') \cap cl(\pi^{-1}(\delta(d_*(p), d_*(p')))$ where $cl(.)$ denotes the closure, there exists a shortest path γ connecting a point x of $\pi^{-1}(p)$ to x' such that $\pi(\gamma) = \delta$ and $d_X(x, x') = d(x') - d(x) = d_*(p') - d_*(p) = d_G(p, p')$.*

Proof. First assume that p' is not a merging point. Let $\gamma_0 : [0, d(x')] \to X$ be any shortest path between r and x' and let γ be the restriction of γ_0 to $[d_*(p), d(x')] = [d_*(p), d_*(p')]$. If the infimum t_0 of the set $I = \{t \in [d_*(p), d_*(p')] : \pi(\gamma(t')) \in \delta, \ \forall t' \geqslant t\}$ is larger than $d_*(p)$, then $\pi(\gamma(t_0))$ then there exists an increasing sequence (t_n) that converges to t_0 such that $\gamma(t_n) \notin \delta$. As a consequence $\delta(t_0)$ is a merging point; a contradiction. So $t_0 = d_*(p)$ and $\gamma(d_*(p))$ intersects $\pi^{-1}(p)$ at a point x.

Now if p' is a merging point, as x' is chosen in the closure of $\pi^{-1}(\delta(d_*(p), d_*(p'))$, for any sufficiently large $n \in \mathbb{N}$ one can consider a sequence of points $x'_n \in \pi^{-1}(\delta(d_*(p') - 1/n))$ that converges to x' and apply the first case to get a sequence of shortest path γ_n from a point $x_n \in \pi^{-1}(p)$ and x'_n. Then applying Arzelà-Ascoli's theorem (see [**?**] 7.5) we can extract from γ_n a sequence of points converging to a shortest path γ between a point $x \in \pi^{-1}(p)$ and x'.

To conclude the proof, notice that from Lemma 3.1 we have $d_G(p, p') = d_*(p') - d_*(p) = d(x') - d(x)$. Since γ is the restriction of a shortest path from r to x we also have $d_X(x, x') = d(x') - d(x)$. □

Notice that from Lemma 3.1, δ is a shortest path and the parametrization by the interval $[d_*(p), d_*(p')]$ can be chosen to be an isometric embedding.

Proposition 4.4. *For any $x, y \in X$ we have*

$$d_X(x, y) \leqslant d_G(\pi(x), \pi(y)) + 2(\beta_1(G) + 1)M$$

where $M = \sup_{p \in G}\{diam(\pi^{-1}(p))\}$ and $\beta_1(G)$ is the first Betti number of G.

Proof. Let δ be a shortest path between $\pi(x)$ and $\pi(y)$. Remark that except at the points $\pi(x)$ and $\pi(y)$ the local maxima of the restriction of d_* to δ are in V_m. Indeed as δ is a shortest path it has to be simple, so if $p \in \delta$ is a local maximum then p has to be a vertex and δ has to pass through two edges having p as end point and pointing to p according to the orientation defined in Section **??**. So p is a merging point. Since δ is simple and V_m is finite, δ can be decomposed in at most $|V_m| + 1$ connected paths along the interior of which the restriction of d_* does not have any local maxima. So along each of these connected paths the restriction of d_* can have at most one local minimum. As a consequence, δ can be decomposed in a finite number of continuous paths $\delta_1, \delta_2, \cdots, \delta_k$ with $k \leqslant 2(|V_m| + 1)$, such that the restriction of d_* to each of these path is strictly monotonic. For any $i \in \{1, \cdots, k\}$ let p_i and p_{i+1} the end points of δ_i with $p_1 = \pi(x)$ and $p_{k+1} = \pi(y)$. We can apply Lemma 4.3 to each δ_i to get a shortest path γ_i in X between a point $x_i \in \pi^{-1}(p_i)$ and a point in $y_{i+1} \in \pi^{-1}(p_{i+1})$ such that $\pi(\gamma_i) = \delta_i$ and $d_X(x_i, y_{i+1}) = d_G(p_i, p_{i+1})$. The sum of the lengths of the paths γ_i is equal to the sum of the lengths of the path δ_i which is itself equal to $d_G(\pi(x), \pi(y))$. Now for any $i \in \{1, \cdots, k\}$, since $\pi(x_i) = \pi(y_i)$

we have $d_X(x_i, y_i) \leqslant M$ and x_i and y_i can be connected by a path of length at most M (x_1 is connected to x and y_{k+1} is connected to y. Gluing these paths to the paths γ_i gives a continuous path from x to y whose length is at most $d_G(\pi(x), \pi(y)) + kM \leqslant d_G(\pi(x), \pi(y)) + 2(|V_m| + 1)M$. Since from Lemma 4.2, $|V_m| \leqslant \beta_1(G)$, we finally get that $d_X(x, y) \leqslant d_G(\pi(x), \pi(y)) + 2(\beta(G) + 1)M$. □

Proposition 4.5. *The map $\pi : X \to G$ is 1-Lipschitz: for any $x, y \in X$ we have*

$$d_G(\pi(x), \pi(y)) \leqslant d_X(x, y).$$

Proof. Let $x, y \in X$ and let $\gamma : I \to X$ be a shortest path from x to y in X where $I \subset \mathbb{R}$ is a closed interval. The path $\pi(\gamma)$ connects $\pi(x)$ and $\pi(y)$ in G.

We first claim that there exists a continuous path Γ contained in $\pi(\gamma)$ connecting $\pi(x)$ and $\pi(y)$ that intersects each vertex of G at most one time. The path Γ can be defined by iteration in the following way. Let $v_1, \cdots v_n \in V$ be the vertices of G that are contained in $\pi(\gamma) \setminus \{\pi(x), \pi(y)\}$ and let $\Gamma_0 = \pi(\gamma) : J_0 = I \to G$. For $i = 1, \cdots n$ let $t_i^- = \inf\{t : \Gamma_{i-1}(t) = v_i\}$ and $t_i^+ = \sup\{t : \Gamma_{i-1}(t) = v_i\}$ and define Γ_i as the restriction of Γ_{i-1} to $J_i = J_{i-1} \setminus (t_i^-, t_i^+)$. The path Γ_i is a connected continuous path (although J_i is a disjoint union of intervals) that intersects the vertices $v_1, v_2, \cdots, v_i$ at most one time. We then define $\Gamma = \Gamma_n : J = J_n \to G$ where $J \subset I$ is a finite union of closed intervals. Notice that Γ is the image by π of the restriction of γ to J and that $\Gamma(t) \in \{v_1, \cdots v_n\}$ only if t is one of the endpoints of the closed intervals defining J.

Now, for each connected component $[t, t']$ of J, $\gamma((t, t'))$ is contained in $\pi^{-1}(e)$ where e is the edge of G containing $\Gamma([t, t'])$. As a consequence,

$$\begin{aligned} d_G(\pi(\gamma)(t), \pi(\gamma)(t')) &= |d_*(\pi(\gamma)(t) - d_*(\pi(\gamma)(t'))| \\ &= |d(\gamma(t)) - d(\gamma(t'))|. \end{aligned}$$

Recalling that $d(\gamma(t)) = d_X(r, \gamma(t))$ and $d(\gamma(t')) = d_X(r, \gamma(t'))$ and using the triangle inequality we get that $|d(\gamma(t)) - d(\gamma(t'))| \leqslant d_X(\gamma(t), \gamma(t'))$. To conclude the proof, since γ is a geodesic path we just need to sum up the previous inequality over all connected components of J:

$$\begin{aligned} d_X(x, y) &\geqslant \sum_{[t,t'] \in cc(J)} d_X(\gamma(t), \gamma(t')) \\ &\geqslant \sum_{[t,t'] \in cc(J)} d_G(\pi(\gamma)(t), \pi(\gamma)(t')) \geqslant d_G(\pi(x), \pi(y)) \end{aligned}$$

where $cc(J)$ is the set of connected components of J. □

The proof of Theorem 4.1 now easily follows from Propositions 4.4 and 4.5.

Proof. (of Theorem 4.1) Consider the set $C = \{(x, \pi(x)) : x \in X\} \subset X \times G$. As π is surjective this is a correspondence between X and G. It follows from Propositions 4.4 and 4.5 that for any $(x, \pi(x)), (y, \pi(y)) \in C$,

$$|d_X(x, y) - d_G(\pi(x), \pi(y))| \leqslant 2(\beta_1(G) + 1)M$$

So C is a $2(\beta_1(G) + 1)M$-correspondence and $d_{GH}(X, G) \leqslant (\beta_1(G) + 1)M$. □

4.2 Bounding M

The two following lemmas, proven in Appendix, allow to bound the diameter of the level sets of π.

Lemma 4.6. *Let (G, d_G) be a connected finite metric graph and let $r \in G$. We denote by $d_r = d_G(r, .) : G \to [0, +\infty)$ the distance to r. For any edge $e \subset G$, the restriction of d_r to e is either strictly monotonic or has only one local maximum. Moreover the length $l = l(e)$ of e is upper bounded by two times the difference between the maximum and the minimum of d_r restricted to e.*

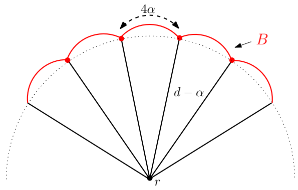


Figure 1: Tightness of the bound in Lemma 4.7: there are 3 edges of length at most 4α and the diameter of B is equal to 20α.

Proof. Let l be the length of E and let $t \mapsto e(t)$, $t \in [0, l]$, be an arc length parametrization of E. Since E is an edge of G, for $t \in [0, l]$ any shortest geodesic γ_t joining r to $e(t)$ must contain either $x_1 = e(0)$ or $x_2 = e(l)$. If it contains x_1 then for any $t' < t$ the restriction of γ_t between r and $e(t')$ is a shortest geodesic containing x_1 and if it contains x_2 then for any $t' > t$ the restriction of γ_t between r and $e(t')$ is a shortest geodesic containing x_2. Moreover in both cases, the function d_r is strictly monotonic along γ. As a consequence, the set $I_1 = \{t \in [0, l] : \text{a shortest geodesic joining } r \text{ to } e(t) \text{ contains } x_1\}$ is a closed interval containing 0. Similarly the set $I_2 = \{t \in [0, l] : \text{a shortest geodesic joining } r \text{ to } e(t) \text{ contains } x_2\}$ is a closed interval containing l and $[0, l] = I_1 \cup I_2$. Moreover d_r is strictly monotonic on $e(I_1)$ and on $e(I_2)$. As a consequence $I_1 \cap I_2$ is reduced to a single point t_0 that has to be the unique local maximum of d_r restricted to E.

The second part of the lemma follows easily from the previous proof: the minimum of d_r restricted to E is attained either at x_1 or x_2 and $d_r(e(t_0)) = d_r(x_1) + t_0 = d_r(x_2) + l - t_0$ is the maximum of d_r restricted to E. We thus obtain that $2t_0 = l + (d_r(x_2) - d_r(x_1))$. As a consequence if $d_r(x_1) \leqslant d_r(x_2)$ then $l/2 \leqslant t_0 = d_r(e(t_0)) - d_r(x_1)$; similarly if $d_r(x_1) \geqslant d_r(x_2)$ then $l/2 \leqslant l - t_0 = d_r(e(t_0)) - d_r(x_2)$. $\square$

Proposition 4.7. *Let (G, d_G) be a connected finite metric graph and let $r \in G$. For $\alpha > 0$ we denote by $N_E(\alpha)$ the number of edges of G of length at most α. For any $d > 0$ and any connected component B of the set $B_{d,\alpha} = \{x \in G : d - \alpha \leqslant d_G(r, x) \leqslant d + \alpha\}$ we have*

$$diam(B) \leqslant 4(2 + N_E(4\alpha))\alpha$$

Proof. Let $x, y \in B$ and let $t \mapsto \gamma(t) \in B$ be a continuous path joining x to y in B. Let E be an edge of G that does not contain x or y and with end points x_1, x_2 such that γ intersects the interior of E. Then $\gamma^{-1}(E)$ is a disjoint union of closed intervals of the form $I = [t, t']$ where $\gamma(t)$ and $\gamma(t')$ belong to the set $\{x_1, x_2\}$. If $\gamma(t) = \gamma(t')$ we can remove the part of γ between t and t' and still get a continuous path between x and y. So without loss of generality we can assume that if γ intersects the interior of E, then E is contained in γ. Using the same argument as previously we can also assume that if γ goes across E, it only does it one time, i.e. $\gamma^{-1}(E)$ is reduced to only one interval. As a consequence, γ can be decomposed in a sequence $[x, v_0], E_1, E_2, \cdot, E_k, [v_k, y]$ where $[x, v_0]$ and $[v_k, y]$ are pieces of edges containing x and y respectively and $E_1 = [v_0, v_1], E_2 = [v_1, v_2]\cdot, E_k = [v_{k-1}, v_k]$ are pairwise distinct edges of G contained in B. It follows from Lemma 4.6 that the lengths of the edges $E_1, \cdots E_k$ and of $[x, v_0]$ and $[v_k, y]$ are upper bounded by 4α. As a consequence the length of γ is upper bounded by $4(k + 2)\alpha$ which is itself upper bounded by $4(N_E(4\alpha) + 2)\alpha$ since the edges $E_1, \cdots E_k$ are pairwise distinct. It follows that $d_G(x, y) \leqslant 4(N_E(4\alpha) + 2)\alpha$. $\square$

The example of the right picture shows that the bound of Lemma 4.7 is tight.

Theorem 4.8. *Let (G, d_G) be a connected finite metric graph and let (X, d_X) be a compact geodesic metric space such that $d_{GH}(X, G) < \varepsilon$ for some $\varepsilon > 0$. Let $x_0 \in X$ be a fixed point and let $d_{x_0} = d_X(x_0, .) : X \to [0, +\infty)$ be the distance function to x_0. Then for $d \geqslant \alpha \geqslant 0$ the diameter of any connected component L of $d_{x_0}^{-1}([d - \alpha, d + \alpha])$ satisfies*

$$diam(L) \leqslant 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon) + \varepsilon$$

where $N_E(4(\alpha + 2\varepsilon))$ is the number of edges of G of length at most $4(\alpha + 2\varepsilon)$. In particular if $\alpha = 0$ and 8ε is smaller that the length of the shortest edge of G then $diam(L) < 17\varepsilon$.

Proof. Let $\varepsilon' > 0$ be such that $d_{GH}(X, G) < \varepsilon' < \varepsilon$. Let $C \subset X \times G$ be an ε'-correspondence between X and G and $(x_0, r) \in C$. we denote by $d_r = d_G(r, .) : G \to [0, +\infty)$ the distance function to r in G. Let $x_a, x_b \in L$ and let $(x_a, y_a), (x_b, y_b) \in C$. There exists a continuous path $\gamma \subseteq L$ joining x_a to x_b. Since C is an ε'-correspondence for any $x \in \gamma$ there exists a point $(x, y) \in C$ such that $d - \alpha - \varepsilon' \leqslant d_r(y) \leqslant d+\alpha+\varepsilon'$. The set of points y obtained in this way is not necessarily a continuous path from y_a to y_b. However one can consider a finite sequence $x_1 = x_a, x_2, \cdots, x_n = x_b$ of points in γ such that for any $i = 1, \cdots n-1$ we have $d_X(x_i, x_{i+1}) < \varepsilon - \varepsilon'$. If $(x_i, y_i) \in C$ then we have $d_G(y_i, y_{i+1}) < \varepsilon - \varepsilon' + \varepsilon' = \varepsilon$. As a consequence, since $d - \alpha - \varepsilon < d - \alpha - \varepsilon' < d_r(y_i) < d + \alpha + \varepsilon' < d + \alpha + \varepsilon$ the shortest geodesic connecting y_i to y_{i+1} in G remains in the set $d_r^{-1}([d - \alpha - 2\varepsilon, d + \alpha + 2\varepsilon])$ and connecting these geodesics for all $i = 1, \cdots, n - 1$ we get a continuous path from y_a to y_b in $d_r^{-1}([d - \alpha - 2\varepsilon, d + \alpha + 2\varepsilon])$. It then follows from Proposition 4.7 that $d_G(y_a, y_b) \leqslant\leqslant 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon)$ and since C is an ε'-correspondence (and so an ε-correspondence), $d_X(x_a, x_b) < 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon) + \varepsilon$. □

From Theorems 4.8 and 4.1 we obtain the following results for the Reeb and α-Reeb graphs.

Theorem 4.9. *Let (X, d_X) be a compact connected path metric space, let $r \in X$ be a fixed base point such that the metric Reeb graph (G, d_G) of the function $d = d_X(r, .) : X \to \mathbb{R}$ is a finite graph. If for a given $\varepsilon > 0$ there exists a finite metric graph $(G', d_{G'})$ such that $d_{GH}(X, G') < \varepsilon$ then we have*

$$d_{GH}(X, G) < (\beta_1(G) + 1)(17 + 8N_{E,G'}(8\varepsilon))\varepsilon$$

where $N_{E,G'}(8\varepsilon)$ is the number of edges of G' of length at most 8ε. In particular if X is at distance less than ε from a metric graph with shortest edge length larger than 8ε then $d_{GH}(X, G) < 17(\beta_1(G) + 1)\varepsilon$.

Theorem 4.10. *Let (X, d_X) be a compact connected path metric space. Let $r \in X$, $\alpha > 0$ and $\mathcal{I}$ be a finite covering of the segment $[0, Diam(X)]$ by open intervals of length at most α such that the α-Reeb graph G_α associated to $\mathcal{I}$ and the function $d = d_X(r, .) : X \to \mathbb{R}$ is a finite graph. If for a given $\varepsilon > 0$ there exists a finite metric graph $(G', d_{G'})$ such that $d_{GH}(X, G') < \varepsilon$ then we have*

$$d_{GH}(X, G_\alpha) < (\beta_1(G_\alpha) + 1)(4(2 + N_{E,G'}(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon) + \varepsilon)$$

where $N_{E,G'}(4(\alpha + 2\varepsilon))$ is the number of edges of G' of length at most $4(\alpha + 2\varepsilon)$. In particular if X is at distance less than ε from a metric graph with shortest edge length larger than $4(\alpha + 2\varepsilon)$ then $d_{GH}(X, G_\alpha) < (\beta_1(G_\alpha) + 1)(8\alpha + 17\varepsilon)$.

5 Recovery of Topology

In this section, we show the following theorem which asserts that the α-Reeb graph G of (X, d) recovers some topology of X.

Theorem 5.1. *Let (X, d_X) be a compact connected path metric space and $(G', d_{G'})$ is a metric graph so that $d_{GH}(X, G') < \varepsilon$. Let $r \in X$, $\alpha > 60\varepsilon$ and $\mathcal{I}\{[0, 2\alpha), (i\alpha, (i+2)\alpha) | 1 \leqslant i \leqslant m\}$ covers the segment $[0, Diam(X)]$ such that the α-Reeb graph G associated to $\mathcal{I}$ and the function $d = d_X(r, .) : X \to \mathbb{R}$ is a finite graph. If no edges of G' is shorter than L and no loops of G' is shorter than $2L$ with $L \geqslant (24\alpha + 9\varepsilon)$, then we have G and G' are homotopy equivalent.*

Since $d_{GH}(X, G') < \varepsilon$, there exists an ε-correspondence between the two spaces, denoted $C(X, G')$. For any subset $V \subset X$, denote $C(V) = \{g' : (x, g') \in C(X, G'), x \in V\}$, and similarly for any subset $U \subset G'$, denote $C(U) = \{x : (x, g') \in C(X, G'), g' \in U\}$. We call $C(V)$ and $C(U)$ are the correspondence set of V and U respectively under $C(X, G')$. Recall that $r \in X$ is the root point. Choose a point $g_r \in C(r)$ and define a distance function $b : G' \to \mathbb{R}$ by $b(g) = d_{G'}(g_r, g)$. Let

$N = \{g_{n_1}, g_{n_2}, \cdots, g_{n_p}\}$ be the vertices of G', i.e., N is the set of vertices whose degree is not equal to two. From the hypothesis of the above theorem, the distance between any pair of vertices g_{n_i}, g_{n_j} with $i \neq j$ is larger than L. For convenience, we also add into the vertices of G' the remaining local maximal/minimal points of the distance function b, which we denote using $M = \{g_{m_1}, \cdots, g_{m_q}\}$. Note any newly added vertex $g_{m_i} \in M$ is of degree two. We call the graph G' before adding the vertices in M the original G', and the edges in the original G' the original edges of G'. An original edge of G' contains at most one vertex in M and thus can be splitted into at most two edges in G'.

Our strategy of proving Theorem 5.1 is to construct some open covers for X and G' and relate the graph G and the grah G' to the nerves of the open covers.

5.1 Construction of open cover for X

We start with the following open cover of X. For each $I_k \in \mathcal{I}$, denote $V_k = d^{-1}(I_k)$. V_k may have several connected components, which can be listed in an arbitrary order. Denote V_k^l the l-th connected component of V_k. Then $\mathcal{V}_0 = \{V_k^l\}_{k,l}$ is an open cover of X. Since at most two elements in $\mathcal{I}$ are overlapped, the nerve of $\mathcal{V}_0$, denoted $N(\mathcal{V}_0)$, is a graph. Moreover by construction $N(\mathcal{V}_0)$ is homotopy equivalent to the alpha-Reeb graph G. The following lemma states that any loop in the nerve $N(\mathcal{V}_0)$ is large, which is useful for the proof of Theorem 5.1. We say an open set $V_{k_1}^{l_1} \in \mathcal{V}_0$ is lower than the open set $V_{k_2}^{l_2} \in \mathcal{V}_0$ if $k_1 < k_2$ and is higher than $V_{k_2}^{l_2}$ if $k_2 > k_1$.

Lemma 5.2. *Let V_k^l and V_j^i are the lowest vertex and the highest vertex of a loop respectively in the nerve $N(\mathcal{V}_0)$. Then under the hypothesis of Theorem 5.1, we have $j - k \geqslant 11$.*

Proof. Let $x_1 \in V_k^l \cap d^{-1}(k\alpha, (k+1)\alpha)$ and $x_2 \in V_j^i \cap d^{-1}((j+1)\alpha, (j+2)\alpha)$. From the hypothesis of the lemma, there are two different paths γ_1, γ_2 connecting x_1 to x_2 so that $\gamma_1 \cap d^{-1}((k+1)\alpha, (j+1)\alpha)$ and $\gamma_2 \cap d^{-1}((k+1)\alpha, (j+1)\alpha)$ are in the different connected components of $d^{-1}((k+1)\alpha, (j+1)\alpha)$. Choose $g_i \in G'$ from $C(x_i)$ for $i = 1, 2$. By a similar construction in the proof of Theorem 4.8, the path γ_i in X for $i = 1, 2$ traces out a path L_i in G' connecting g_1 to g_2 so that L_i lies in $b^{-1}(k\alpha - 2\epsilon, (j+2)\alpha + 2\epsilon)$. One can verify that L_1 and L_2 are different and thus form a loop in G', denoted β. We have $b(\beta) \in (k\alpha - 2\epsilon, (j+2)\alpha + 2\epsilon)$. We claim the range of any loop, in particular β, under the function b, covers an interval with the length at least $\frac{L}{2}$. If the claim holds, then we have $(j - k + 2)\alpha + 4\varepsilon \geqslant \frac{L}{2}$, which implies $j - k \geqslant 11$ from the hypothesis of Theorem 5.1. Indeed, if β contains at least two vertices in N, then it is obvious that the range of β under the function b covers an interval the length at least $\frac{L}{2}$ as any edge in G' (before adding the vertices in M) is longer than L. Now consider the case where β contains one vertex in N, say g_a. If β does not contain g_r, then there is exactly one local maximum on β, say g_b. If β contains g_r, let $g_b = g_r$. The removal of g_a and g_b cuts β into two pieces. Along either piece, the function b is monotonic. As the length of β is longer than $2L$. We have $b(\beta)$ covers an interval with length longer than L. If β contains g_r, Finally, if β contains no vertex in N, then G' is a single loop β and the claim obviously holds. $\square$

In the following, we modify this open cover by merging while preserve the homotopy type of its nerve. The main purpose of the merging operation is make it easy to relate the open cover of X to the open cover of G' constructed in Section 5.2. The merging operation is done in two steps.

For any vertex $g \in M \cup N$ of G', we construct a connected open set $V(g)$ as the union of a subset of the open cover $\mathcal{V}_0$ as follows. If $b(g) \geqslant \frac{\alpha}{2}$, then there exists a unique positive integer k' s.t. $k'\frac{\alpha}{2} \leqslant b(g) < (k'+1)\frac{\alpha}{2}$. Let $k = [\frac{k'+1}{2}] - 1 \geqslant 0$, and one can verify that $(k + \frac{1}{2})\alpha \leqslant b(g) \leqslant (k + \frac{3}{2})\alpha$. Therefore for all $x \in C(g)$, $d(x) \in [(k + \frac{1}{2})\alpha - \varepsilon, (k + \frac{3}{2})\alpha + \varepsilon] \subset I_k$. Moreover $C(g)$ is contained in $V_k^l \subset V_k$ for some l. Indeed, if not, assume $x_1, x_2 \in C(g)$ with $x_i \in V_k^i$ for $i \in \{1, 2\}$. By the definition of V_k^i, the geodesic connecting x_1 and x_2 must pass through a point x_0 outside of V_k, which means $d_X(x_i, x_0) \geqslant |d(x_i) - d(x_0)| \geqslant \frac{\alpha}{2} - \epsilon$. Then $d_X(x_1, x_2) \geqslant \alpha - 2\epsilon$ which contradicts to the fact that $d_X(x_1, x_2) \leqslant d_{G'}(g, g) + \varepsilon \leqslant \varepsilon$. Now we construct the open set $V(g)$ as the union of the elements in the open cover $\mathcal{V}_0$ having non-empty intersection with V_k^l, i.e.,

$$V(g) = \cup_{V \in \mathcal{V}_0 \text{ and } V \cap V_k^l \neq \emptyset} V.$$

In the case where $b(g) < \frac{\alpha}{2}$, we construct the open set $V(g) = V_0 \cup V_1 = d^{-1}([0, 3\alpha))$. Note in both cases, $V(g)$ is a connected open set of X. We abuse the notation and also denote $V(g)$ the subset of $\mathcal{V}_0$ whose union is the open set $V(g)$. What $V(g)$ represents will be clear from the context. For convenience, we call the V_k^l containing $C(g)$ the center of $V(g)$. Note that it is possible that $V(g) = V(g')$ for two different vertices g, g'.

Now we obtain an intermediate open cover of X

$$\mathcal{V} = \{V(g) : g \in M \cup N\} \cup \{V \in \mathcal{V}_0 : V \notin V(g) \forall g \in M \cup N\}$$

Note as a set, $\mathcal{V}$ does not have duplicated elements, i.e., if $V(g) = V(g')$ for $g \neq g'$, then $\mathcal{V}$ only contains one copy of $V(g)$. We call an open set $V(g) \in \mathcal{V}$ for any $g \in M \cup N$ critical and the remaining ones regular. The following two lemmas describe the properties of both critical open sets and the regular open sets.

Lemma 5.3. *Under the hypothesis of Theorem 5.1, we have for any vertex $g \in M \cup N$,*
i) $d(V(g)) \subset [s\alpha, (s+4)\alpha]$ for some integer $s \geqslant 0$, and
ii) for any point $x \in \cup_{V \in \mathcal{V}_0 \setminus V(g)} V$ and any $g_x \in C(x) \subset G'$, $d_{G'}(g, g_x) \geqslant \frac{\alpha}{2} - 2\varepsilon$.

Proof. The claim (i) is obvious from the construction of $V(g)$. We now prove claim (ii). In the case where $b(g) < \frac{\alpha}{2}$, for any $x \in \cup_{V \in \mathcal{V}_0 \setminus V(g)} V$, one have $d(x) > 2\alpha$ and $b(g_x) > 2\alpha - \varepsilon$. Thus $d_{G'}(g, g_x) \geqslant |b(g_x) - b(g)| > 2\alpha - \varepsilon - \frac{\alpha}{2} > \frac{\alpha}{2} - 2\varepsilon$. Now consider the case where $b(g) \geqslant \frac{\alpha}{2}$. If $d(x) \notin I_k$, then $d_X(x, y) \geqslant \frac{\alpha}{2} - \varepsilon$ for any point $y \in C(g)$ from the construction of $V(g)$, which implies $d_{G'}(g_x, g) \geqslant \frac{\alpha}{2} - 2\varepsilon$. Otherwise $d(x) \in I_k$. Then x is not in V_k^l and the geodesic from x to any point $y \in C(g)$ must pass $x_0 \notin V_k$. This implies that $d_X(x, y) > d_X(x_0, y) \geqslant \frac{\alpha}{2} - \epsilon$ and $d_{G'}(g_x, g) \geqslant d_X(x, y) - \epsilon \geqslant \frac{\alpha}{2} - 2\epsilon$. This proves the lemma. □

Lemma 5.4. *For any regular open set $V \in \mathcal{V}$, V is also an open set in $\mathcal{V}_0$. Moreover, it is of degree two in the nerve of $N(\mathcal{V}_0)$ with one neighbhoring vertex higher than V and one neighboring vertex lower than V.*

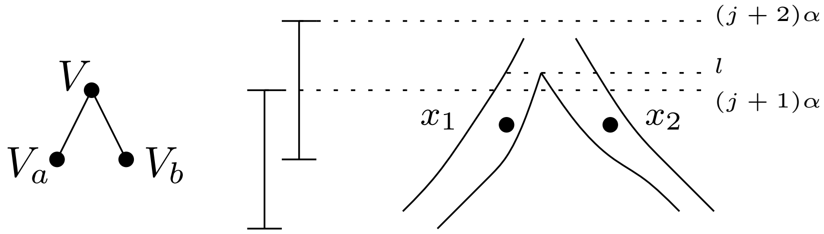


Figure 2: V with two lower neighbourhoods.

Proof. We prove the lemma by contradiction. Assume $V \in \mathcal{V}_0 \setminus \cup_{g \in M \cup N} V(g)$ has two neighboring vertices, say V_a, V_b, which are lower than V. Without loss of generality, assume $d(V) \subset I_j$ and $d(V_a)$ and $d(V_b)$ are subsets of I_{j-1}. Let $x_a \in V_a$ and $x_b \in V_b$ such that $(j-1)\alpha < d(x_a), d(x_b) < j\alpha$. As V_a and V_b both have non-empty intersection with V, there exist a path in $d^{-1}((j-1)\alpha, (j+2)\alpha)$. Now let $l = inf\{s :$ there exists a path connecting x_a, x_b in $d^{-1}((j-1)\alpha, s] \cap (V \cup V_a \cup V_b)\}$. We have $l \geqslant j\alpha$ as V_a, V_b are disconnected.

We prove the lemma by contradiction. Assume $V \in \mathcal{V} / \cup_{g \in M \cup N} V(g)$ has more than one lower neighbourhood, say, V_a, V_b. Without loss of generality, let $d(V) \subset I_j$, then $d(V_a)$ and $d(V_b)$ are I_{j-1}. Let $x_a \in V_a$ and $x_b \in V_b$ such that $(j-1)\alpha < d(x_a), d(x_b) < (j - \frac{1}{2})\alpha$, as V_a and V_b both have non-empty intersection with V, there exist a path in $d^{-1}((j-1)\alpha, (j+2)\alpha)$. Now let $l = inf\{$ there is a path connecting x_a, x_b in $d^{-1}((j-1)\alpha, l] \cap (V \cup V_a \cup V_b)\}$, obviously $l \geqslant j\alpha$ otherwise V_a, V_b are connected to each other. Specially we can find two points x_1, x_2 in the different connected components of $d^{-1}((j-1)\alpha, l)\}$ such that $d(x_1) = d(x_2) = l - 3.5\epsilon$. Therefore x_1, x_2 are path connected in

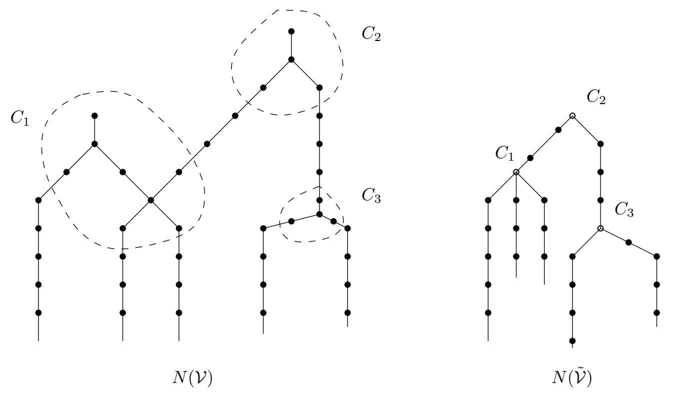


Figure 3: $N(\mathcal{V})$ and $N(\tilde{\mathcal{V}})$.

$d^{-1}([l-3.5\epsilon, l])$, on the other hand, the diameter of this component is upper bounded by 31ϵ, i.e. $d_X(x_1, x_2) \leqslant 31\epsilon$. Obviously $d_X(x_1, x_2) \geqslant 2(l-(l-3.5\epsilon)) = 7\epsilon$. Let $(x_i, g_i) \in C(X, G')$, then $b(g_i) \in [l-4.5\epsilon, l+2\epsilon]$ and $6\epsilon \leqslant d_X(x_1, x_2) - \epsilon \leqslant d_{G'}(g_1, g_2) \leqslant d_X(x_1, x_2) + \epsilon \leqslant 32\epsilon$. If the geodesic contains no node or local maximal point, $d_{G'}(g_1, g_2) = |b(g_1) - b(g_2)| \leqslant 2\epsilon < 6\epsilon$: impossible. If the geodesic contains more than one node of G', then the length of the geodesic exceeds L, which contradicts to $d_{G'}(g_1, g_2) \leqslant 32\epsilon$, thus there is at most one node in the geodesic.

Since we can find a path connecting x_1 and x_2 in $d^{-1}([l-3.5\epsilon, l])$, there exists a path connecting g_1 and g_2 in $b^{-1}([l-5.5\epsilon, l+2\epsilon])$. Now we claim that the geodesic between g_1 and g_2 must be contained in the corresponding path in G': otherwise the path together with the geodesic is a loop in G', so the length of the loop is larger than $2L - d_{G'}(g_1, g_2) \geqslant 2L - 32\epsilon$, therefore the range of the path in G' is at least $\frac{2L-32\epsilon}{2} = L - 16\epsilon > 7.5\epsilon$: contradiction. As mentioned before, there must be some critical point(s) in the geodesic, thus for all the critical point(s) g_c, $b(g_c) \in [l-5.5\epsilon, l+2\epsilon]$.

Another observation about the critical point in the geodesic is we can at least find a g_c for one of g_1, g_2, say, g_1, such that $d_{G'}(g_c, g_1) = b(g_c) - b(g_1)$. In fact if there is no node on the geodesic, then g_1, g_2 are in the same edge of G', so there is one and only one critical point, more precisely, local maximal point g_m in that geodesic. Obviously $g_c = g_m$ and $d_{G'}(g_c, g_i) = b(g_c) - b(g_i), i = 1, 2$. Another possible case is there is exactly one node g_n in the geodesic, then g_1 and g_2 are in two distinct edges which are connected by g_n respectively. If there is no local maximal point between g_i and g_n, then $b(g_n)$ must be larger than $b(g_i)$, otherwise the geodesic for g_1, g_2 lies in $b^{-1}([l-5.5\epsilon, l-2.5\epsilon])$, then we can find a path in X connecting x_1, x_2 in $d^{-1}([l-7.5\epsilon, l-0.5\epsilon])$: contradiction, thus $g_c = g_n$. At last if there is a local maximal point g_m between one of (or both) paths from g_i to g_n, without loss of generality, from g_1 to g_n, then $d_{G'}(g_m, g_1) = b(g_m) - b(g_1)$, i.e. $g_c = g_m$.

WLOG we assume we find g_c a critical point in the geodesic between g_1, g_2 such that $d_{G'}(g_c, g_1) = b(g_c) - b(g_1)$, noticing that the geodesic is contained in $b^{-1}([l-5.5\epsilon, l+2\epsilon])$, thus $b(g_c) \leqslant l + 2\epsilon$ and $d_{G'}(g_c, g_1) \leqslant l + 2\epsilon - (l-4.5\epsilon) = 6.5\epsilon$. On the other hand, we can find a point x_0 in $V \cap d^{-1}(l)$ such that there exists an x_1' satisfies $d_X(x_0, x_1') = d(x_0) - d(x_1') = 3.5\epsilon$. Also we have $d_X(x_1, x_1') \leqslant 17\epsilon$, then $d_X(x_1, x_0) \leqslant 20.5\epsilon$ and $d_{G'}(g_0, g_1) \leqslant 21.5\epsilon$. Finally we got a point $x_0 \in V$ such that $d_{G'}(g_0, g_c) \leqslant 6.5\epsilon + 21.5\epsilon = 28\epsilon$, however according to the definition of V, for any critical point g_c, $d_{G'}(g_0, g_c) \geqslant \frac{\alpha}{2} - 2\epsilon > 28\epsilon$: contradiction.

Using almost the same argument we can also prove V can not have more than one higher neighbourhood. At last if V is of degree 1, then obviously $C(V)$ is close to some local maximal point, which is also an impossible situation. $\square$

We now perform a second step of merging. Two critical open set $V(g_1)$ and $V(g_2)$ in $N(\mathcal{V})$ are said to be close if there is a path in the 1-skeleton of the nerve $N(\mathcal{V})$ connecting $V(g_1)$ and $V(g_2)$ and passing through at most one more open set in $\mathcal{V}$. If exists, such open set is said to connect the critical open sets $V(g_1)$ and $V(g_2)$. We have the following properties for two critical open sets being close.

Lemma 5.5. *i. For any two vertices $g_{n_1}, g_{n_2} \in N$, $V(g_{n_1})$ and $V(g_{n_2})$ can not be close;*

ii. *If $V(g_{m_1})$ and $V(g_{m_2})$ are close for any two vertices $g_{m_1}, g_{m_2} \in M$, then there must exist a vertex $g_n \in N$ such that $V(g_{m_1})$ and $V(g_{m_2})$ are both close to $V(g_n)$;*

iii. *For any $g_m \in M$, there exists at most one $g_n \in N$ such that $V(g_m)$ and $V(g_n)$ are close.*

Proof. First we consider two critical points g_1, g_2 of G' satisfy $d_{G'}(g_1, g_2) = |b(g_1) - b(g_2)|$, i.e. the geodesic connecting them is in fact a monotonous path. Now we claim that if $d_{G'}(g_1, g_2) > 5\alpha + 4\epsilon$, then $V(g_1)$ and $V(g_2)$ are not close. WLOG, assume that $b(g_1) > b(g_2)$, let the range of the center of $V(g_1)/V(g_2)$ is subset to I_j/I_k, one has $b(g_1) \in [(j+0.5)\alpha, (j+1.5)\alpha]$, $d(x_1) \in [(j+0.5)\alpha - \epsilon, (j+1.5)\alpha+\epsilon]$(similar for g_2), now $d_{G'}(g_1, g_2) = b(g_1) - b(g_2) \leqslant d(x_1) + \epsilon - (d(x_2) - \epsilon) = d(x_1) - d(x_2) + 2\epsilon$, on the other hand, $d(x_1) - d(x_2) \leqslant (j + 1.5)\alpha + \epsilon - ((k + 0.5)\alpha - \epsilon) = (j - k + 1)\alpha + 2\epsilon$, so $(j-k+1)\alpha + 4\epsilon > 5\alpha + 4\epsilon$, meaning $j - k > 4$ thus there must be at least two V_r between the centres of $V(g_1), V(g_2)$.

Remark 5.6. It seems natural that the shortest path in $N(\mathcal{V})$ from $V(g_1)$ and $V(g_2)$ are connected via V_r as the path between g_1, g_2 are monotonous. I believe it is true but am not sure about the proof. That is also the obstacle I mentioned before. And **if we have that guarantee**, then for any two critical points, we separate the geodesic at the interior critical points and therefore we have a series of monotonous segments of the geodesic, now if one of the segments is longer than $5\alpha + 4\epsilon$, then the path contains at least two V_r, meaning that $V(endpoints)$ are not close.

The only possible case that g_1 and g_2 are connected via some other critical point(s), and all of the segments are not longer than $5\alpha + 4\epsilon$ is g_1, g_2 are two local maximal points connected via a node g_n.

i)By assumption, $d_{G'}(g_{n_1}, g_{n_2}) > L > 10\alpha + 9\alpha$, $V(g_{n_1})$ and $V(g_{n_2})$ are not close;

ii)The geodesic between two local maximal points must contain some node, if not, then the two local maximal points are in the same edge of G', then the restriction of $b(g)$ in this edge has at most one local maximum, that means $g_{m_1} = g_{m_2}$. Furthermore, if the geodesic passes more than one node, then $d_{G'} g_{m_1}, g_{m_2} > L > 10\alpha + 9\epsilon$, contradicts to $V(g_{m_1}), V(g_{m_2})$ are close. Therefore there is only one node in the geodesic, denote it by g_n. In fact we only need to prove that the path connecting the centres of $V(g_{m_1})$ and $V(g_{m_2})$ must pass the centre of $V(g_n)$. Assume intervals I_j, I_{k_1}, I_{k_2} well-contains $b(g_n), b(g_{m_1}), b(g_{m_2})$:

case 1, if $V(g_{m_1}), V(g_{m_2})$ are neither not close to $V(g_n)$, that is $k_i \geqslant j+6$, specially, if $max\{k_1, k_2\}$ is smaller than $j+10$, then we are going to find a loop in $d^{-1}(j\alpha, (j+12)\alpha)$, that contradicts to lemma **??**. Thus we have $k_1 + k_2 \geqslant j + 6 + j + 11 = 2j + 17$, then $d_{G'}(g_{m_1}, g_{m_2}) \geqslant (k_1 + k_2 - 2j - 2)\alpha = 15\alpha > 10\alpha + 9\epsilon$: impossible.

case 2, if $V(g_{m_1})$ is close to $V(g_n)$ while $V(g_{m_2})$ is not. Still we need $k_2 \geqslant j + 11$, at the same time $k_1 \leqslant j + 5$ as the closeness. Then $k_2 - k_1 \geqslant 6$, contradicts to the assumption $V(g_{m_1})$ and $V(g_{m_2})$ are close.

iii)If $V(g_m)$ is close to $V(g_{n_1})$ and $V(g_{n_2})$ at the same time, then $d_{G'}(g_{n_1}, g_{n_2}) \leqslant d_{G'}(g_m, g_{n_1}) + d_{G'}(g_m, g_{n_2}) \leqslant 20\alpha + 18\epsilon < L$, contradicts to the assumption. So any $V(g_m)$ can at most close to one $V(g_n)$.

First we claim that if $d_{G'}(g_1, g_2) > 10\alpha + 9\epsilon$, then $V(g_1)$ and $V(g_2)$ are not close. Since $d_{G'}(g_1, g_2) \leqslant 10\alpha + 13\epsilon$, the range of any path connecting g_1 and g_2 is at least $5\alpha + 6.5\epsilon$. On the other hand, if $V(g_1)$ is close to $V(g_2)$, then we have the $C(g_1), C(g_2)$ are connected in $d^{-1}((k + \frac{1}{2})\alpha - \epsilon, (k + 4 + \frac{3}{2})\alpha + \epsilon)$. That implies we can find a path from g_1 to g_2 in $b^{-1}((k + \frac{1}{2})\alpha - 3\epsilon, (k + 4 + \frac{3}{2})\alpha + 3\epsilon)$, thus the range of the path is at most $5\alpha + 6\epsilon$, thus it is not the case.

i)By assumption, $d_{G'}(g_{n_1}, g_{n_2}) > L > 10\alpha + 9\alpha$, $V(g_{n_1})$ and $V(g_{n_2})$ are not close;

ii)The geodesic between two local maximal points must contain some node, if not, then the two local maximal points are in the same edge of G', then the restriction of $b(g)$ in this edge has at most one local maximum, that means $g_{m_1} = g_{m_2}$. Furthermore, if the geodesic passes more than one node, then

$d_{G'}g_{m_1}, g_{m_2} > L > 10\alpha + 9\epsilon$, contradicts to $V(g_{m_1}), V(g_{m_2})$ are close. Therefore there is only one node in the geodesic, denote it by g_n. In fact we only need to prove that the path connecting the centres of $V(g_{m_1})$ and $V(g_{m_2})$ must pass the centre of $V(g_n)$. Assume intervals I_j, I_{k_1}, I_{k_2} well-contains $b(g_n), b(g_{m_1}), b(g_{m_2})$:

case 1, if $V(g_{m_1}), V(g_{m_2})$ are neither not close to $V(g_n)$, that is $k_i \geqslant j+6$, specially, if $max\{k_1, k_2\}$ is smaller than $j+10$, then we are going to find a loop in $d^{-1}(j\alpha, (j+12)\alpha)$, that contradicts to lemma **??**. Thus we have $k_1 + k_2 \geqslant j + 6 + j + 11 = 2j + 17$, then $d_{G'}(g_{m_1}, g_{m_2}) \geqslant (k_1 + k_2 - 2j - 2)\alpha = 15\alpha > 10\alpha + 9\epsilon$: impossible.

case 2, if $V(g_{m_1})$ is close to $V(g_n)$ while $V(g_{m_2})$ is not. Still we need $k_2 \geqslant j + 11$, at the same time $k_1 \leqslant j + 5$ as the closeness. Then $k_2 - k_1 \geqslant 6$, contradicts to the assumption $V(g_{m_1})$ and $V(g_{m_2})$ are close.

iii)If $V(g_m)$ is close to $V(g_{n_1})$ and $V(g_{n_2})$ at the same time, then $d_{G'}(g_{n_1}, g_{n_2}) \leqslant d_{G'}(g_m, g_{n_1}) + d_{G'}(g_m, g_{n_2}) \leqslant 20\alpha + 18\epsilon < L$, contradicts to the assumption. So any $V(g_m)$ can at most close to one $V(g_n)$. □

Now we are ready to further merge the open sets in $\mathcal{V}$ to obtain the final open cover $\tilde{\mathcal{V}}$ of X as follows. For any vertex $g_n \in N$ of G', Let $\tilde{V}(g_n)$ be the subset of $\mathcal{V}$ consisting of (1) $V(g_n)$, and (2) any critical open set $V(g) \in \mathcal{V}$ for any $g \in M$ which is close to $V(g_n)$, and (3) any regular open set $V \in \mathcal{V}$ connecting $V(g_n)$ and some critical open set. We abuse the notation and also denote $\tilde{V}(g_n)$ the open set of the union of the open sets in $\tilde{V}(g_n)$. What $\tilde{V}(g_n)$ represents will be clear from the context.

Let $\tilde{\mathcal{V}}_N = \{V \in \mathcal{V} : V \in \tilde{V}(g_n) \text{ for some } g_n \in N\}$. The open cover $\tilde{\mathcal{V}} = \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2 \cup \tilde{\mathcal{V}}_3$ of X consists of three types of open sets:

(1) $\tilde{\mathcal{V}}_1 = \{\tilde{V}(g_n) : g_n \in N\}$;

(2) $\tilde{\mathcal{V}}_2 = \{V(g) : g \in M \text{ and } V(g) \notin \tilde{\mathcal{V}}_N\}$

(3) $\tilde{\mathcal{V}}_3 = \{V \in \mathcal{V} : V \text{ is regular and } V \notin \tilde{\mathcal{V}}_N\}$.

We summarize the properties for the open cover $\tilde{\mathcal{V}}$ in the following corollary, which follows from Lemma 5.3, Lemma 5.4, and Lemma 5.5.

Corollary 5.7. *The open sets in $\tilde{\mathcal{V}}$ satisfy the following properties.*

- *$\tilde{V}(g_1)$ and $\tilde{V}(g_2)$ are disjoint for two different $g_1, g_2 \in N$.*
- *For any two open sets $V_1, V_2 \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$, any path in the 1-skeleton of the nerve $N(\tilde{\mathcal{V}})$ connecting V_1, V_2 consists of at least two elements from $\tilde{\mathcal{V}}_3$.*
- *Any open set $V \in \tilde{\mathcal{V}}_3$ is also an open set in $\mathcal{V}_0$. It is of degree two in the 1-skeleton of the nerve $N(\tilde{\mathcal{V}})$ with one neighboring vertex lower than V and one neighboring vertex higher than V. Moreover any point $g \in C(V) \subset G'$ is at least $\frac{\alpha}{2} - 2\epsilon$ away from any vertex of G'.*

Proposition 5.8. *Under the hypothesis of Theorem 5.1, $N(\tilde{\mathcal{V}})$ and $N(\mathcal{V}_0)$ are homotopy equivalent.*

Proof. Figure 3 shows an example of $N(\mathcal{V}_0)$ and $N(\tilde{\mathcal{V}})$. We obtain the open covering $\tilde{\mathcal{V}}$ by merging some particular open sets in $\mathcal{V}_0$.

It suffices to prove that all sub-complex of form C_i is a tree in $N(\mathcal{V})$. If $C_i = V(g_c), g_c \in M \cup N$, than the range of $d(x)$ restricted in C_i is at most 4α. Otherwise C_i is obtained by combining all the $V(g_m)'s$ that close to some $V(g_n)$. Assume $I_1 = (k\alpha, (k+2)\alpha), I_2 = (j\alpha, (j+2)\alpha)$ well contain $V(g_n)$ and $V(g_m)$ respectively(i.e. $b(g_n)$ and $b(g_m)$ are in the middle part of the two intervals), if I_2 is higher than I_1, since they are close, $j - k \leqslant 4$. Other if I_2 is lower than I_1, it is impossible $k - j > 1$, otherwise $b(g_n) \geqslant (k + \frac{1}{2})\alpha \geqslant (j + 2 + \frac{1}{2}\alpha) \geqslant (j + \frac{3}{2})\alpha + \alpha \geqslant b(g_m) + \alpha$, thus $b(g_n) \geqslant b(g_m) + \alpha > b(g_m)$. That means the geodesic from g_n to g_m must contains some other node

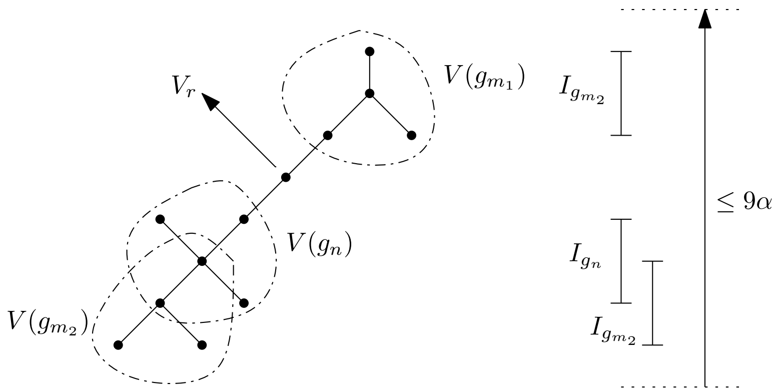


Figure 4: The upper bound of the range of C_i.

in G', therefore $d_{G'}(g_n, g_m) > L$, $V(g_n)$ and $V(g_m)$ can not be close. To conclude, the range of $d(x)$ restricted to C_i is at most 9α. According to lemma **??**, it contains no loop in $N(\mathcal{V})$, so it is a tree in $N(\mathcal{V})$.

Since $N(\tilde{\mathcal{V}})$ is obtained by collapsing the sub-complex C_i of $N(\mathcal{V})$. Without loss of generality, we can collapse one C_i at one time, following the proposition **??**, $N(\mathcal{V}) \simeq N(\mathcal{V})/C_i$, the number of C_i is finite, so at the end we get $N(\mathcal{V}) \simeq N(\tilde{\mathcal{V}})$.

□

5.2 Construction of open cover for G'

In this section, we construct an open cover G' based on the open cover $\tilde{\mathcal{V}}$ of X. For a open set $V \in \mathcal{V}_0$, we construct a connected open set $U_V \in G'$ so that $C(V) \in U_V$ as follows. Let $l = \min\{d(V)\}$ and $u = \max\{d(V)\}$. We have $u - l \leqslant 2\alpha$. Let $\overline{U} = b^{-1}([u + 2\varepsilon, l - 2\varepsilon])$, and then $C(V) \subset \overline{U}$. Since $u - l + 4\varepsilon < 2\alpha + 4\varepsilon < \frac{L}{4}$, one can verify that there is no loop in $\overline{U}$ and thus $\overline{U}$ consists of a set of trees. We claim $C(V)$ is contained in one of the trees. Indeed, for any two $g_1, g_2 \in C(V)$, we have $l - \varepsilon < b(g_1), b(g_2) < u + \varepsilon$. Now let $x_i \in V$ so that $g_i \in C(x_i)$ for $i = 1, 2$. Let γ be a path in V connecting x_1 and x_2. Using an argument in the proof of Theorem 4.8, γ can trace out a path in $\overline{U}$ connecting g_1 and g_2, which implies that $C(V)$ is contained in a tree in $\overline{U}$. Let U_V denote that tree. Let $\mathcal{U}_0 = \{U_V : V \in \mathcal{V}_0\}$. It is obvious that $\mathcal{U}_0$ is an open cover of G'. We now merge the elements in $\mathcal{U}_0$ to construct a new open cover according to the way in which the elements in $\mathcal{V}_0$ are merged to obtain $\tilde{\mathcal{V}}$. Specifically, from our construction of $\tilde{\mathcal{V}}$, any open set $\tilde{V} \in \tilde{\mathcal{V}}$ is the union of a subset of open sets of $\mathcal{V}_0$. We also denote this subset using $\tilde{V}$. Let $U_{\tilde{V}} = \{U_V : V \in \tilde{V} \subset \mathcal{V}_0\}$. We also denote $U_{\tilde{V}}$ is the open set of the union of open sets in $U_{\tilde{V}}$.

Consider an open set $\tilde{V} \in \tilde{\mathcal{V}}_3$. As it is also a regular open set in $\mathcal{V}$ and thus an open set in $\mathcal{V}_0$, $d(\tilde{V})$ = $(p\alpha, (p+2)\alpha)$ for some integer $p > 0$. From Corollary 5.7, any point in $C(\tilde{V})$ is at least $\frac{\alpha}{2} - 2\epsilon$ away from any vertex in $M \cup N$ and any point in $U_{\tilde{V}}$ is at least $\frac{\alpha}{2} - 4\epsilon$ away from any vertex in $M \cup N$. Thus $U_{\tilde{V}}$ is a segment in G' without any branches. We shrink $U_{\tilde{V}}$ to obtain a new open set $\tilde{U}_{\tilde{V}} = U_{\tilde{V}} \cap b^{-1}(p\alpha + 2\epsilon, (p+2)\alpha - 2\varepsilon)$. For any open set $\tilde{V} \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$, let $\tilde{U}_{\tilde{V}} = U_{\tilde{V}}$. Thus we obtain

$$\tilde{\mathcal{U}} = \{\tilde{U}_{\tilde{V}} : \tilde{V} \in \tilde{\mathcal{V}}\}.$$

One can verify that $\tilde{\mathcal{U}}$ is an open cover of G'. Moreover we have the following two lemmas which relate the nerve $N(\tilde{\mathcal{V}})$ to G'.

Proposition 5.9. *Under the hypothesis of Theorem 5.1, the nerve $N(\tilde{\mathcal{V}})$ and the nerve $N(\tilde{\mathcal{U}})$ are isomorphic as graph.*

Proof. It suffices to prove the following three claims.

- Claim (i): For any two $\tilde{V}_i, \tilde{V}_j \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$, $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

  Any path in $N(\tilde{\mathcal{V}})$ connecting $\tilde{V}_i$ and $\tilde{V}_j$ must pass through at least two open sets in $\tilde{\mathcal{V}}_3$, which are regular open sets in $\mathcal{V}$. From Lemma **??**, any regular set has two neighbors in the nerve $N(\mathcal{V})$ one lower and one higher, WLOG, assume $\tilde{V}_i$ is higher than $\tilde{V}_j$. We have $\inf\{d(x) : x \in \tilde{V}_i\} \geqslant \alpha + sup\{d(x)|x \in \tilde{V}_i\}$, which implies $\inf\{b(g)|g \in \tilde{U}_{\tilde{V}_i}\} \geqslant \alpha + sup\{b(g)|g \in \tilde{U}_{\tilde{V}_j}\} - 2\epsilon > sup\{b(g)|g \in \tilde{U}_{\tilde{V}_j}\}$. Thus $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

- Claim (ii): For any two $\tilde{V}_i, \tilde{V}_j \in \tilde{\mathcal{V}}_3$, $\tilde{V}_i \cap \tilde{V}_j = \emptyset$ if and only if $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

  If $\tilde{V}_i \cap \tilde{V}_j \neq \emptyset$, assume $\tilde{V}_i$ is the only neighboring vertex in the nerve $N(\mathcal{V})$ higher than $\tilde{V}_j$. Let $d(\tilde{V}_j) = (p\alpha, (p+2)\alpha$ and $d(\tilde{V}_i) = ((p+1)\alpha, (p+3)\alpha)$. Choose a point x from $\tilde{V}_i \cap \tilde{V}_j$ so that $d(x) = (p + \frac{3}{2})\alpha)$. We have $C(x) \in \tilde{U}_{\tilde{V}_i} cap \tilde{U}_{\tilde{V}_j}$, which shows $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} \neq \emptyset$.

  If $\tilde{V}_i \cap \tilde{V}_j = \emptyset$. Let $d(\tilde{V}_i) = (p\alpha, (p+2)\alpha)$ and $d(\tilde{V}_i) = (q\alpha, (q+2)\alpha)$. If $|p - q| \geqslant 2$, it is obvious that $\tilde{U}_{\tilde{V}_i} cap \tilde{U}_{\tilde{V}_j} = \emptyset$. Now assume that $q - p \leqslant 1$, which forces the shortest path connecting $\tilde{V}_i$ and $\tilde{V}_j$ in $N(\tilde{\mathcal{V}})$ must pass through some open set $\tilde{V} \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$. By Lemma **??**, $d_{G'}(C(\tilde{V}_i), g) \geqslant \frac{\alpha}{2} - 2\varepsilon$ and $d_{G'}(C(\tilde{V}_j), g) \geqslant \frac{\alpha}{2} - 2\varepsilon$ for any vertex $g \in M \cup N$ such that $V(g) \in \tilde{V}$. Thus $d_{G'}(C(\tilde{V}_i), C(\tilde{V}_j)) \geqslant \alpha - 4\varepsilon$, which implies $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

- Claim (iii): For any $\tilde{V}_i \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$ and any $\tilde{V}_j \in \tilde{\mathcal{V}}_3$, $\tilde{V}_i \cap \tilde{V}_j = \emptyset$ if and only if $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

  First assume that $\tilde{V}_i$ and $\tilde{V}_j$ have a non-empty intersection. As $\tilde{V}_j \in \tilde{\mathcal{V}}_3$, it is a regular open set in (V) which has one higher neighering vertex and one lower neighboring vertex in $N((V))$. Let $d(\tilde{V}_j) = (p\alpha, (p+2)\alpha)$ for some integer $p > 0$. $\tilde{V}_i$ is a subset of open sets in $\mathcal{V}_0$. Let $V \in \tilde{V}_i$ be the open set in $\mathcal{V}_0$ so that $V \cap \tilde{V}_j \neq \emptyset$. WLOG, assume $d(V) \subset ((p+1)\alpha, (p+3)\alpha)$. In fact, we claim $d(V) = ((p+1)\alpha, (p+3)\alpha)$. If not, one can verify that V is the center of an open set $V(g) \in \mathcal{V}$ from some vertex g of G'. This is impossible as $\tilde{V}_j$ being a regular open set can not intersect with the center of $V(g)$ for any g. we choose a point in $x \in \tilde{V}_j \cap V$ so that $d(x) = ((p+2)\alpha - 4\epsilon)$. Since $b(C(x)) \subset ((p+2)\alpha - 5\epsilon, (p+2)\alpha - 3\epsilon)$, $C(x) \in \tilde{U}_{\tilde{V}_j} \cap \tilde{U}_{\tilde{V}_i}$ and thus $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} \neq \emptyset$.

  Second assume $\tilde{V}_i \cap \tilde{V}_j = \emptyset$. If any path in the nerve $N(\tilde{\mathcal{V}})$ connecting $\tilde{V}_i$ and $\tilde{V}_j$ passes through some open set in $\tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$, then we are done based on Claim (i). Now assume there is a path β in the nerve $N(\tilde{\mathcal{V}})$ connecting $\tilde{V}_i$ and $\tilde{V}_j$ only passing through open sets in $\tilde{\mathcal{V}}_3$. Since any open set in $\tilde{\mathcal{V}}_3$ is a regular set in $\mathcal{V}$, the worst scenery is that β contains no intermediate open sets. In this worst scenery, due to the shrinking operation on $\tilde{U}_{\tilde{V}_j}$, one can verify that $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$.

□

Proposition 5.10. *Under the hypothesis of Theorem 5.1, $N(\tilde{\mathcal{U}})$ is homotopy equivalent to G'.*

Proof. As we have proved, $\tilde{\mathcal{U}}$ is an open covering of G'. Since any edge on the original G' has a length longer than L, one can verify that any element of $\tilde{\mathcal{U}}$ contains no loop and thus is a tree, and in particular is contratible. Furthermore, the union of any two elements of $\tilde{\mathcal{U}}$ does not contains a loop. This means that if two elements of $\tilde{\mathcal{U}}$ intersect with each other, their intersection is connected and thus contractible. Following from Nerve lemma, we have $N(\tilde{\mathcal{U}})$ is homotopy equivalent to G'. □

Proof of Theorem 5.1. The theorem follows from Proposition 5.8, Proposition 5.9, Proposition 5.10 and the fact that the nerve $N(\mathcal{V}_0)$ is homotopy equivalent to the α-Reeb graph G.

6 Algorithm

In this section, we describe an algorithm for computing α-Reeb graph for some $\alpha > 0$. We assume the input of the algorithm includes a neighboring graph $H = (V, E)$, a function $l : E \to \mathbb{R}^+$ specifying the

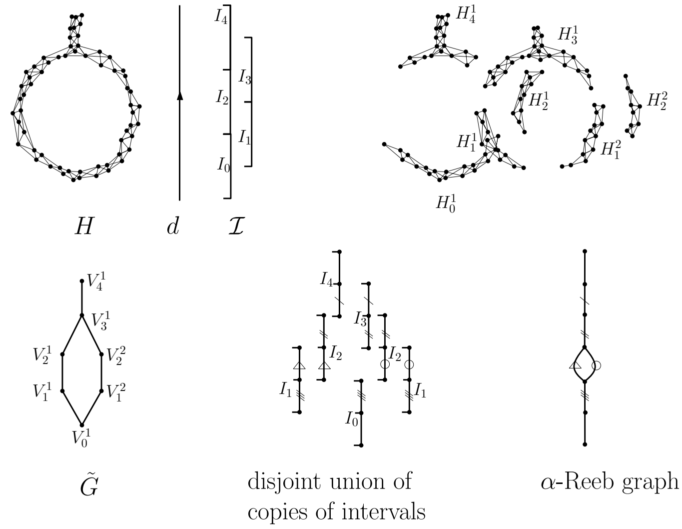


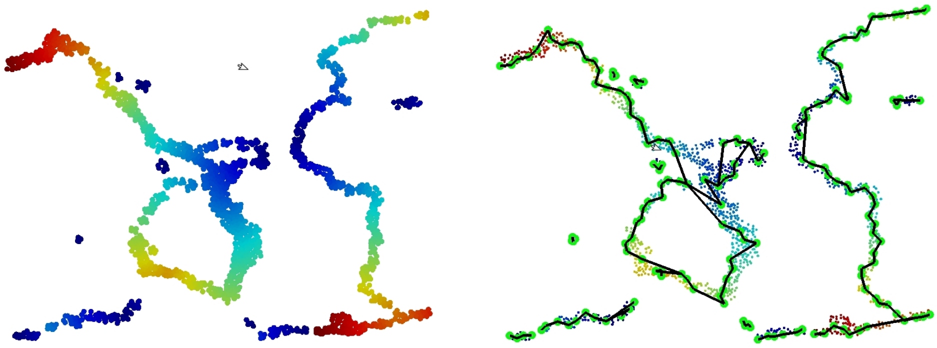
Figure 5: Illustration of the different steps of the algorithm for computing α-Reeb graph. In the disjoint union of copies of intervals, the subintervals marked with same labels are identified in the α-Reeb graph.

edge length and a parameter α. In the applications where the input is given as a set of points together with pairwise distances, i.e., a finite metric space, one can generate the neighboring graph H as a Rips graph of the input points with the parameter chosen as a fraction of α. We assume H is connected as one can apply the algorithm to each connected component otherwise.

Our algorithm, whose different steps are illustrated on Figure 5, can be described as follows. In the first step, we fix a node of H as the root r and then obtain the distance function $d : V \to \mathbb{R}^+$ by computing $d(v)$ as the graph distance from the node v to r. In the second step, we apply the Mapper algorithm [**?**] to the nodes V with filter d to construct a graph $\tilde{G}$. Specifically, let $\mathcal{I} = \{(i\alpha, (i+1)\alpha), ((i+0.5)\alpha, (i+1.5)\alpha) | 0 \leqslant i \leqslant m\}$ so that $\cup_{I_k \in \mathcal{I}} I_k$ covers the range of the function d. We say an interval $I_{k_1} \in \mathcal{I}$ is lower than another interval $I_{k_2} \in \mathcal{I}$ if the midpoint of I_{k_1} is smaller than that of I_{k_2}. Now let H_k be the subgraph of H restricted to $V_k = d^{-1}(I_k)$. Namely two nodes in H_k are connected with an edge if they are in H. Notice that each subgraph H_k may have several connected components, which can be listed in an arbitrary order. Denote H_k^l the l-th connected component of H_k and V_k^l its set of nodes. Then of $\{V_k^l\}_{k,l}$ is a cover of V and the graph $\tilde{G}$ constructed by the Mapper algorithm is the 1-skeleton of the nerve of that cover. Namely, each node in $\tilde{G}$ represents an element in $\{V_k^l\}_{k,l}$, i.e., a subset of nodes in V. Two nodes $V_{k_1}^{l_1}$ and $V_{k_2}^{l_2}$ are connected with an edge if $V_{k_1}^{l_1} \cap V_{k_2}^{l_2} \neq \emptyset$.

In the final step, we represent each node V_k^l in $\tilde{G}$ using a copy of the interval I_k. As mentioned in the Section **??**, α-Reeb graph is a quotient space of the disjoint union of those copies of intervals. Specifically, for an edge in $\tilde{G}$, let $V_{k_1}^{l_1}$ and $V_{k_2}^{l_2}$ be its endpoints. Then I_{k_1} and I_{k_2} must be partially overlapped. We identify the overlap part of these two intervals. After identifying the overlapped intervals for all edges in $\tilde{G}$, the resulting quotient space is the α-Reeb graph. Algorithmically, the identification is performed as follows. We split each copy of internal I_k into two by adding a point in the middle. Now think of it as a graph with two edges and label one of them upper and the other lower. Notice that two overlapped intervals I_{k_1} and I_{k_2} can not be exactly the same. One must be lower than the other. To identify their overlapped part, we identify the upper edge of the lower interval with the lower edge of the upper interval.

The time complexity of the above algorithm is dominated by the computation of the distance function in the first step, which is $O(|E| + |V| \log |V|)$. The computation of the connected components in the second step is $O(|V| \log |V|)$ based on union-find data structure. In the final step, there are at most $O(|V|)$ number of the copies of the intervals. Based on union-find data structure, the identification can also be performed in $O(|V| \log |V|)$ time.



(a) (b)

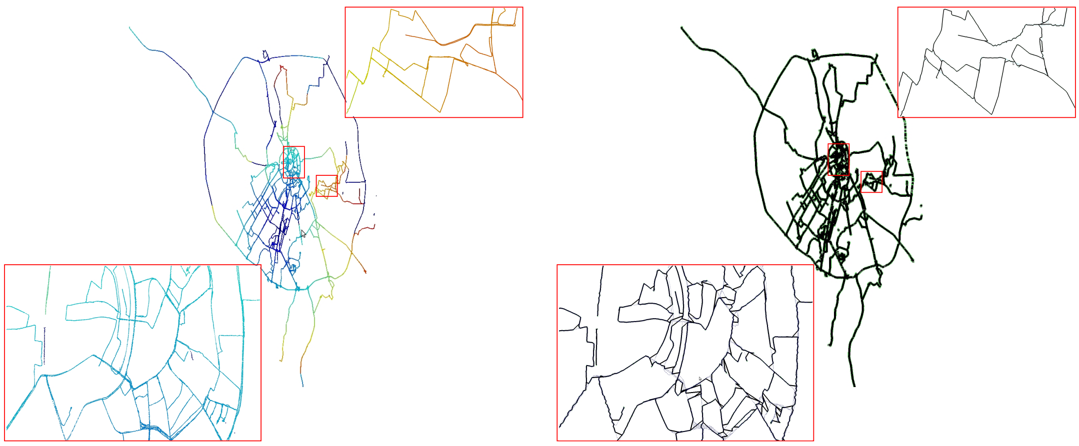
Figure 6: (a) The distance functions d on each connected components. The value increases from cold to warm colors. (b) The reconstructed α-Reeb graph.

7 Experiments

In this section, we illustrate the performances of our algorithm on three different data sets. The first data set was obtained from USGS Earthquake Search [**?**]. It consists of earthquakes epicenters locations collected, between 01/01/1970 and 01/01/2010, in the rectangular area between latitudes -75 degrees and 75 degrees and longitude -170 degrees and 10 degrees, and of magnitude greater than 5.0. This raw earthquake data set contains the coordinates of the epicenters of 12790 earthquakes that are mainly located around geological faults. We follow the procedure described in [**?**] to remove outliers and randomly sampled 1600 landmarks. Finally, we computed a neighboring graph from these landmarks with parameter 4. The length of an edge in this graph is the Euclidean distance between its endpoints. For each connected component, we fix a root point and compute the graph distance function d to the root point as shown in Figure 6(a). We also set α equals 4 and apply our algorithm to the above data to obtain the α-Reeb graph. In general α-Reeb graph is an abstract metric graph. In this example, for the purpose of visualization, we use the coordinates of the landmarks to embed the graph into the plane as follows. Recall that for a copy of interval I_k representing the node V_k^l in $\tilde{G}$, we split it into two by adding a point in the middle. We embed the endpoints of the interval to the landmarks of the minimum and the maximum of the function d in V_k^l, and the point in the middle to the landmark of the median of the function d in V_k^l. Figure 6(b) shows the embedding of the α-Reeb graph. Note this embedding may introduce metric distortion, i.e., the Euclidean length of the edge may not reflect the length of the corresponding edge in the α-Reeb graph.

The second data set is that of 500 GPS traces tagged “Moscow” from OpenStreetMap [**?**]. Since cars move on roads, we expect the locations of cars to provide information about the metric graph structure of the Moscow road network. We first selected a metric ϵ-net on the raw GPS locations with $\epsilon = 0.0001$ using furthest point sampling. Then, we computed a neighboring graph from the samples with parameter 0.0004. Again for each connected component, we fix a root point and compute the graph distance function d to the root point as shown in Figure 7(a). Set α also equals 0.0004 and compute the α-Reeb graph. Again, we use the same method as above to embed the α-Reeb graph into the plane, as shown in Figure 7(b).

To evaluate the quality of our α-Reeb graph for each data set, we computed both original pairwise distances, and pairwise distances approximated from the constructed α-Reeb graph. For GPS traces, we randomly select 100 points as the data set is too big to compute all pairwise distances. We also evaluated the use of α-Reeb graph to speed up distance computations by showing reductions in computation time. Only pairs of points in the same connected component are included because we obtain zero error for the pairs of vertices that are not. Statistics for the size of the reconstructed graph, error of approximate distances, and reduction in computation time are given in Table 1.



(a) (b)

Figure 7: (a) The distance functions d on each connected components. The value increases from cold to warm colors. (b) The reconstructed α-Reeb graph.

	#OP	#OE	#N	#E	GRT	ODT	ADT	Mean	Median
GPS traces	82541	313415	21644	21554	46.8	15.27	0.96	6.5%	5.3%
Earthquake	1600	26996	147	137	0.32	1.12	0.01	14.1%	12.5%

Table 1: #OP (#OE, #N, #E) stands for the number of original points (original edges, nodes, edges in α-Reeb graph). The graph reconstruction time (GRT) is the total time of computing distance function and reconstructing the graph. The original (ODT), respectively approximate (ADT), distance computation time shows the total time of computing these distances using the original, respectively reconstructed, graph. All times are in seconds. The last two columns show the mean and median metric distortions.

The third data set we consider is also obtained from GPS traces. Roads are often split so that cars in different directions run in different lanes. In particular, this is the true for highways. In addition, when two roads cross in GPS coordinates, they may bypass through a tunnel or an evaluated bridge and thus the road network itself may not cross. Such directional information is contained in the GPS traces. We encode this directional information by stacking several consecutive GPS coordinates to form a point in a higher dimensional space. In this way, we obtain a new set of points in this higher dimension space. Then we build a neighboring graph for this new set of points based on L_2 norm and apply our algorithm to recover the road network. In particular, although the paths intersect at the cross in GPS coordinates, the road network does not and this should be detected by our algorithm. To test the above strategy, we extract those GPS traces from the above "Moscow" dataset which pass through a highway crossing as shown in Figure 8(a). Since GPS records the position based on time, we resample the traces so that the distances between any two consecutive samples is the same among all traces. Then we apply the above algorithm to the resampled traces. Figure 8(c) and (d) show the reconstructed graph which recovers the road network of this highway crossing.

8 Discussion

We have proposed a method to approximate path metric spaces using metric graphs with bounded Gromov-Hausdorff distortion, and illustrated the performances of our method on a few data sets. Here we point out a few possible directions for future work. First, notice that the α-Reeb graph is a quotient space where the quotient map is 1-Lipschitz and thus the metric only gets contracted. In addition, the distance from a point to the chosen root is exactly preserved. Therefore, one always reduces the metric

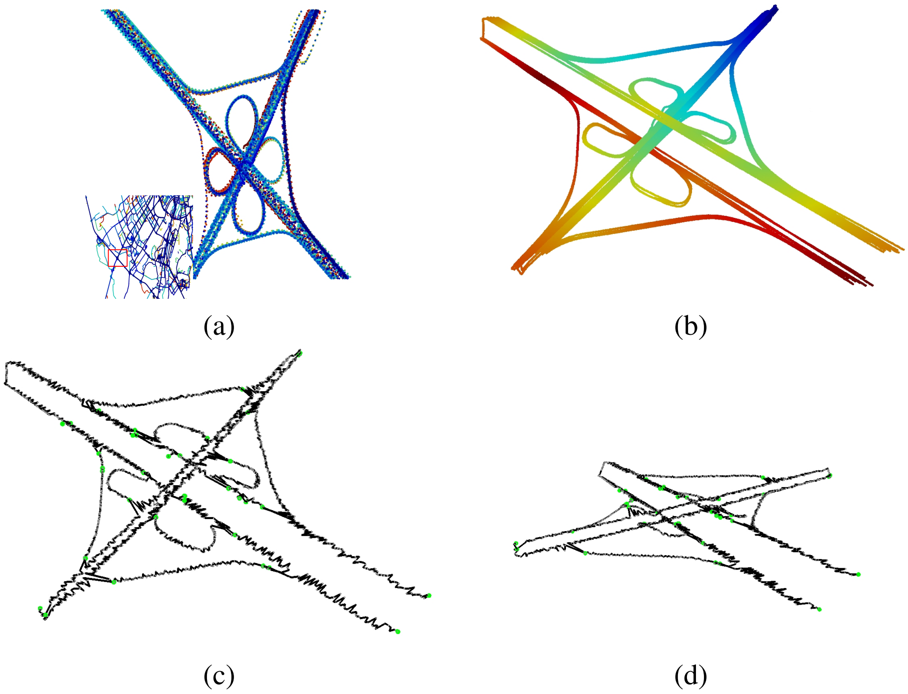


Figure 8: (a) GPS traces passing through a highway crossing in Moscow . (b) The distance function. (c) and (d)The reconstructed α-Reeb graph viewed from two perspectives.

distortion by taking the maximum of the graph metrics of different root points. It is interesting to study the strategy of sampling root points to obtain the smallest metric distortion with the fixed number of root points. Second, our method is sensitive to the noise. One can preprocess the data and remove the noise and then apply our algorithm. Nevertheless, it is interesting to see if the algorithm can be improved to handle noise.

Acknowledgments

The authors acknowledge Daniel Müllner and G. Carlsson for fruitful discussions and for providing code for the Mapper algorithm. They acknowledge the European project CG-Learning EC contract No. 255827; the ANR project GIGA (ANR-09-BLAN-0331-01); The National Basic Research Program of China (973 Program 2012CB825501); The NSF of China (11271011).