

# Gromov–Hausdorff Approximation of Filamentary Structures Using Reeb-Type Graphs

Frédéric Chazal<sup>1</sup> · Ruqi Huang<sup>1</sup> · Jian Sun<sup>2</sup>

Received: 1 August 2014 / Revised: 15 January 2015 / Accepted: 11 February 2015 /

Published online: 14 April 2015

© Springer Science+Business Media New York 2015

**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 performance of our algorithm on a few datasets.

**Keywords** Reeb graph · Gromov-Hausdorff distance · Topological data analysis · Metric spaces

**Mathematics Subject Classification** 68U05 · 51F99

---

Editors in charge: Siu-Wing Cheng and Olivier Devillers

---

Preliminary version in *Proceedings of the 30th Annual Symposium on Computational Geometry, 2014*.

---

Frédéric Chazal  
frederic.chazal@inria.fr

Ruqi Huang  
ruqi.huang@inria.fr

Jian Sun  
jsun@math.tsinghua.edu.cn

<sup>1</sup> INRIA Saclay, Palaiseau, France

<sup>2</sup> Tsinghua University, Beijing, China

# 1 Introduction

## 1.1 Motivation

With the advance of sensor technology, computing power and the 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, these data are not embedded in Euclidean spaces and come as (finite) sets of points with pairwise distance 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 [7,31,35]. 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, [9,11,22,37]) and new tools such as the Mapper algorithm [34] have given rise to new algorithms to extract and visualize geometric and topological information from metric data without the need of an embedding into a 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.

## 1.2 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 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  $(G', d_{G'})$  in both geometry and topology. Our approach relies on the notion of Reeb graph (and some variants of it introduced in Sect. 3) and our main theoretical results are stated in the following two theorems.

**Theorem 3** (Recovery of Geometry) *Let  $(X, d_X)$  be a compact connected geodesic space, and let  $r \in X$  be a fixed base point such that the metric Reeb graph  $(G, d_G)$  of*

the function  $d = d_X(r, \cdot) : X \rightarrow \mathbb{R}$  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  $8\varepsilon$  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  $\varepsilon$  from a metric graph with shortest edge larger than  $8\varepsilon$ , then  $d_{GH}(X, G) < 34(\beta_1(G) + 1)\varepsilon$ .

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

**Theorem 5** (Recovery of Topology) *Let  $(X, d_X)$  be a compact connected path metric space and  $(G', d_{G'})$  be a metric graph so that  $d_{GH}(X, G') < \varepsilon$ . Let  $r \in X$ ,  $\alpha > 60\varepsilon$ , and  $\mathcal{I} = \{[0, 2\alpha), (i\alpha, (i+2)\alpha) \mid 1 \leq i \leq m\}$  covers the segment  $[0, \text{Diam}(X)]$  such that the  $2\alpha$ -Reeb graph  $G$  associated to  $\mathcal{I}$  and the function  $d = d_X(r, \cdot) : X \rightarrow \mathbb{R}$  is a finite graph. If no edges of  $G'$  are shorter than  $L$  and no loops of  $G'$  are shorter than  $2L$  with  $L \geq 32\alpha + 9\varepsilon$ , then  $G$  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 points (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 Sect. 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 [29, 33] 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  $\alpha$ -Reeb graph, inspired from [34] and related to the recently introduced notion of graph induced complex [17], that is easier to compute than the Reeb graph but also comes with approximation guarantees (see Theorem 4). As a consequence, our algorithm runs in almost linear time (see Sect. 6).

### 1.3 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. [4] use multiscale anisotropic strips to detect linear structure, while Genovese et al. [25,27] and more recently [26] 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 focussing on the global geometric structure of the filaments network.

In computational geometry, the curve reconstruction problem from points sampled on a curve in a euclidean space has been extensively studied and several efficient algorithms have been proposed [3,15,16]. Unfortunately, these methods restrict 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, Aanjaneya et al. [1] 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 [1], our algorithm not only comes with metric guarantees but also whatever the input data are, 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 [24], which 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 [24] also always outputs a graph but it does not come with metric guarantees. Recently, Bauer et al. [6] define a metric based on the function for Reeb graph and show that it is stable under Gromov–Hausdorff distance. The implementation of our algorithm relies on the Mapper algorithm [34], which provides a way to visualize datasets 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 provide 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 [2,5,13] or distribution of trees [18,23] 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 2). Finally, we remark that the recovery of filament structure is also studied in various applied settings, including road networks [12,36], galaxies distributions [14].

Part of the result (Theorem 3) shown in the paper also appears in [10]. The paper is organized as follows: The basic notions and definitions used throughout the paper are recalled in Sect. 2. The Reeb and  $\alpha$ -Reeb graphs endowed with a natural metric are introduced in Sect. 3, and the approximation results in metric are stated and proven in

Sect. 4, and the results of recovery of topology are stated and proven in Sect. 5. Our algorithm is described in Sect. 6 and experimental results are presented and discussed in Sect. 7.

## 2 Preliminaries

Recall that a metric space is a pair  $(X, d_X)$  where  $X$  is a set and  $d_X : X \times X \rightarrow \mathbb{R}$  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  $d_X(x, z) \leq d_X(x, y) + d_X(y, z)$ . Two compact spaces  $(X, d_X)$  and  $(Y, d_Y)$  are isometric if there exists a bijection  $\phi : X \rightarrow Y$  that preserves the distances, namely for any  $x, x' \in X$ ,  $d_Y(\phi(x), \phi(x')) = d_X(x, x')$ . 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 ([8, Def. 7.3.17]):

**Definition 1** Let  $(X, d_X)$  and  $(Y, d_Y)$  be two compact metric spaces. Given  $\varepsilon > 0$ , an  $\varepsilon$ -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  $(x, y), (x', y') \in C$ ,  $|d_X(x, x') - d_Y(y, y')| \leq \varepsilon$ .

**Definition 2** The *Gromov–Hausdorff distance* between two compact metric spaces  $(X, d_X)$  and  $(Y, d_Y)$  is defined by

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

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.<sup>1</sup> In the sequel of the paper, we consider compact path metric spaces. It follows from the Hopf–Rinow theorem (see [28, 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.<sup>2</sup> A continuous path  $\delta : I \rightarrow X$ , where  $I$  is a real interval or the unit circle is said to be *simple* if it is not self-intersecting, i.e., if  $\delta$  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 [8, Sect. 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.

<sup>1</sup> See [28, Chap. 1] for the definition of the length of a continuous curve in a general metric space.

<sup>2</sup> Recall that a minimizing geodesic in  $X$  is any curve  $\gamma : I \rightarrow X$ , where  $I$  is a real interval, such that  $d_X(\gamma(t), \gamma(t')) = |t - t'|$  for any  $t, t' \in I$ .

### 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  $d : X \rightarrow \mathbb{R}$  be the distance function to  $r$ , i.e.,  $d(x) = d_X(r, x)$ .

#### 3.1 The Reeb Graph

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

#### 3.2 The $\alpha$ -Reeb Graph

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 properties very similar to the Reeb graph. Let  $\alpha > 0$  and let  $\mathcal{I} = \{I_i\}$  be a covering of the range of  $d$  by open intervals of length at most  $\alpha$ . 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  $\alpha$ -Reeb graph<sup>3</sup> of  $d$  and we denote by  $\pi : X \rightarrow G_\alpha$  the quotient map. Notice that  $\pi$  is continuous and as  $X$  is path connected,  $G_\alpha$  is path connected. The function  $d$  induces a function  $d_* : G_\alpha \rightarrow \mathbb{R}_+$  that satisfies  $d = d_* \circ \pi$ . The relation defined by: for any  $g, g' \in G_\alpha$ ,  $g \leq_{G_\alpha} g'$  if and only if  $d_*(g) \leq d_*(g')$  and there exists a continuous path  $\gamma$  in  $G_\alpha$  connecting  $g$  to  $g'$  such that  $d \circ \gamma$  is non-decreasing, and making  $G_\alpha$  a partially ordered set.

The  $\alpha$ -Reeb graph is closely related to the graph constructed by the Mapper algorithm introduced in [34], making its computation much easier than the Reeb graph (see Sect. 6).

Notice that without making assumptions on  $X$  and  $d$ , in general  $G$  and  $G_\alpha$  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  $d^{-1}(I_i)$ ,  $I_i \in \mathcal{I}$ , is finite, the  $\alpha$ -Reeb graph also turns out to be a finite directed acyclic graph. This happens in most applications, for example,

<sup>3</sup> Strictly speaking we should call it the  $\alpha$ -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.

when  $(X, d_X)$  is a finite simplicial complex or a compact semi-algebraic (or more generally a compact subanalytic space) with  $d$  being semi-algebraic (or subanalytic).

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

### 3.3 A Metric on Reeb and $\alpha$ -Reeb Graphs

Let us 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  $\gamma$  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 1** *If  $\delta$  is a path joining two points  $p, p' \in G$  such that  $d_* \circ \delta$  is strictly increasing, then  $\delta$  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  $\delta$  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,  $\delta$  cannot go through a given edge more than one time. As a consequence,  $\delta$  can be decomposed in a finite sequence of pieces  $e_0 = [p, p_1], e_1 = [p_1, p_2], \dots, e_{n-1} = [p_{n-1}, p_n], e_n = [p_n, p']$ , where  $e_0$  and  $e_n$  are the segments joining  $p$  and  $p'$  to one of the endpoints of the edges that contain them and  $e_1, \dots, e_{n-1}$  are edges. So, the length of  $\delta$  is equal to  $(d_*(p_1) - d_*(p)) + (d_*(p_2) - d_*(p_1)) + \dots + (d_*(p') - d_*(p_n)) = d_*(p') - d_*(p)$  and  $d_G(p, p') \leq d_*(p') - d_*(p)$ .

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



## 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  $\pi$ . An upper bound of  $M$  is given in the next section.

**Theorem 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  $M = \sup_{p \in G} \{\text{diam}(\pi^{-1}(p))\}$  is the supremum of the diameters of in the level sets of  $\pi$ .

Remark that as  $\beta_1(G) \leq \beta_1(X)$ , from the above theorem,  $d_{GH}(X, G)$  is upper bounded by the quantities depending only on the input  $X$ . The proof of Theorem 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 Sect. 3. 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 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 persistence homology theory [21]. First remark that, as  $\pi(r)$  is the only local minimum of  $d_*$ , the sublevel sets of the function  $d_* : G \rightarrow \mathbb{R}_+$  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  $\pi$  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  $\alpha$  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  $|V_m| \leq \beta_1(G)$ .  $\square$

The following lemma shows 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 1 below.

**Lemma 3** *Let  $p, p' \in G$  and let  $\delta : [d_*(p), d_*(p')] \rightarrow G$  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  $x' \in \pi^{-1}(p') \cap \text{cl}(\pi^{-1}(\delta(d_*(p), d_*(p'))))$  where  $\text{cl}(\cdot)$  denotes the closure, there exists a shortest path  $\gamma$  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  $\gamma_0 : [0, d(x')] \rightarrow X$  be any shortest path between  $r$  and  $x'$  and let  $\gamma$  be the restriction of  $\gamma_0$  to  $[d_*(p), d(x')] =$



$[d_*(p), d_*(p')]$ . If the infimum  $t_0$  of the set  $I = \{t \in [d_*(p), d_*(p')] : \pi(\gamma(t')) \in \delta, \forall t' \geq t\}$  is larger than  $d_*(p)$ , then there exists an increasing sequence  $(t_n)$  that converges to  $t_0$  such that  $\pi(\gamma(t_n)) \notin \delta$ . 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  $\gamma_n$  from a point  $x_n \in \pi^{-1}(p)$  and  $x'_n$ . Then applying Arzelà–Ascoli's theorem (see [19, 7.5]), we can extract from  $\gamma_n$  a sequence of points converging to a shortest path  $\gamma$  between a point  $x \in \pi^{-1}(p)$  and  $x'$ .

To conclude the proof, notice that from Lemma 1 we have  $d_G(p, p') = d_*(p') - d_*(p) = d(x') - d(x)$ . Since  $\gamma$  is the restriction of a shortest path from  $r$  to  $x$ , we also have  $d_X(x, x') = d(x') - d(x)$ .  $\square$

Notice that from Lemma 1,  $\delta$  is a shortest path and the parametrization by the interval  $[d_*(p), d_*(p')]$  can be chosen to be an isometric embedding.

**Proposition 1** *For any  $x, y \in X$  we have*

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

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

*Proof* Let  $\delta$  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  $\delta$  are in  $V_m$ . Indeed as  $\delta$  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  $\delta$  has to pass through two edges having  $p$  as end point and pointing to  $p$  according to the orientation defined in Sect. 3. So  $p$  is a merging point.

Since  $\delta$  is simple and  $V_m$  is finite,  $\delta$  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,  $\delta$  can be decomposed in a finite number of continuous paths  $\delta_1, \delta_2, \dots, \delta_k$  with  $k \leq 2(|V_m| + 1)$  such that the restriction of  $d_*$  to each of these paths is strictly monotonic. For any  $i \in \{1, \dots, k\}$ , let  $p_i$  and  $p_{i+1}$  be the end points of  $\delta_i$  with  $p_1 = \pi(x)$  and  $p_{k+1} = \pi(y)$ . We can apply Lemma 3 to each  $\delta_i$  to get a shortest path  $\gamma_i$  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  $\gamma_i$  is equal to the sum of the lengths of the path  $\delta_i$  which is itself equal to  $d_G(\pi(x), \pi(y))$ . Now for any  $i \in \{1, \dots, k\}$ , since  $\pi(x_i) = \pi(y_i)$ , we have  $d_X(x_i, y_i) \leq M$  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  $\gamma_i$  gives a continuous path from  $x$  to  $y$  whose length is at most  $d_G(\pi(x), \pi(y)) + kM \leq d_G(\pi(x), \pi(y)) + 2(|V_m| + 1)M$ . Since from Lemma 2,  $|V_m| \leq \beta_1(G)$ , we finally get that  $d_X(x, y) \leq d_G(\pi(x), \pi(y)) + 2(\beta(G) + 1)M$ .  $\square$

**Proposition 2** *The map  $\pi : X \rightarrow G$  is 1-Lipschitz: for any  $x, y \in X$  we have*

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

*Proof* Let  $x, y \in X$  and let  $\gamma : I \rightarrow X$  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  $\Gamma$  contained in  $\pi(\gamma)$  connecting  $\pi(x)$  and  $\pi(y)$  that intersects each vertex of  $G$  at most one time. The path  $\Gamma$  can be defined by iteration in the following way: Let  $v_1, \dots, v_n \in V$  be the vertices of  $G$  that are contained in  $\pi(\gamma) \setminus \{\pi(x), \pi(y)\}$  and let  $\Gamma_0 = \pi(\gamma) : J_0 \rightarrow G$ ,  $J_0 = I$ . For  $i = 1, \dots, n$ , let  $t_i^- = \inf\{t : \Gamma_{i-1}(t) = v_i\}$  and  $t_i^+ = \sup\{t : \Gamma_{i-1}(t) = v_i\}$  and define  $\Gamma_i$  as the restriction of  $\Gamma_{i-1}$  to  $J_i = J_{i-1} \setminus (t_i^-, t_i^+)$ . The path  $\Gamma_i$  is a connected continuous path (although  $J_i$  is a disjoint union of intervals) that intersects the vertices  $v_1, v_2, \dots, v_i$  at most one time. We then define  $\Gamma = \Gamma_n : J = J_n \rightarrow G$ , where  $J \subset I$  is a finite union of closed intervals. Notice that  $\Gamma$  is the image by  $\pi$  of the restriction of  $\gamma$  to  $J$  and that  $\Gamma(t) \in \{v_1, \dots, v_n\}$  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  $|d(\gamma(t)) - d(\gamma(t'))| \leq d_X(\gamma(t), \gamma(t'))$ . To conclude the proof, since  $\gamma$  is a geodesic path, we just need to sum up the previous inequality over all connected components of  $J$

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

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

The proof of Theorem 1 now easily follows from Propositions 1 and 2.

*Proof* (of Theorem 1) Consider the set  $C = \{(x, \pi(x)) : x \in X\} \subset X \times G$ . As  $\pi$  is surjective, there is a correspondence between  $X$  and  $G$ . It follows from Propositions 1 and 2 that for any  $(x, \pi(x)), (y, \pi(y)) \in C$ ,

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

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

## 4.2 Bounding the Diameter $M$

The following two lemmas allow us to bound  $M$ , the diameter of the level sets of  $\pi$ .

**Lemma 4** *Let  $(G, d_G)$  be a connected finite metric graph and let  $r \in G$ . We denote by  $d_r = d_G(r, \cdot) : G \rightarrow [0, +\infty)$  the distance to  $r$ . For any edge  $e \subset G$ , the restriction of  $d_r$  to  $e$  either is 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  $e$ .*

*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  $\gamma_t$  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  $\gamma_t$  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  $\gamma_t$  between  $r$  and  $e(t')$  is a shortest geodesic containing  $x_2$ . Moreover in both cases, the function  $d_r$  is strictly monotonic along  $\gamma$ . 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  $E$ .

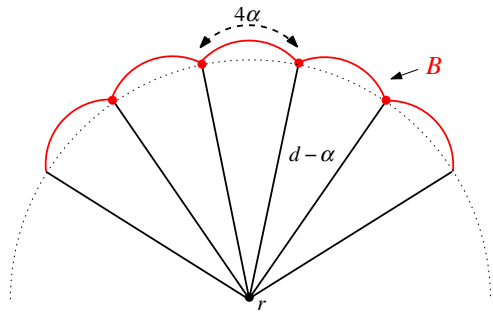
The second part of the lemma follows easily from the previous proof: the minimum of  $d_r$  restricted to  $E$  is attained either at  $x_1$  or  $x_2$  and  $d_r(e(t_0)) = d_r(x_1) + t_0 = d_r(x_2) + l - t_0$  is the maximum of  $d_r$  restricted to  $E$ . We thus obtain that  $2t_0 = l + (d_r(x_2)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

**Proposition 3** *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  $\alpha$ . For any  $d > 0$  and any connected component  $B$  of the set  $B_{d,\alpha} = \{x \in G : d - \alpha \leq d_G(r, x) \leq d + \alpha\}$ , we have*

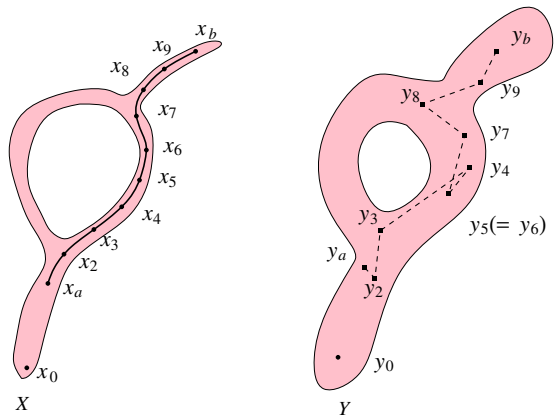
$$\text{diam}(B) \leq 4(2 + N_E(4\alpha))\alpha.$$

*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  $\gamma$  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  $\gamma$  between  $t$  and  $t'$  and still get a continuous path between  $x$  and  $y$ . So without loss of generality, we can assume that if  $\gamma$  intersects the interior of  $E$ , then  $E$  is contained in  $\gamma$ . Using the same argument we can also assume that if  $\gamma$  goes across  $E$ , it only does it one time, i.e.,  $\gamma^{-1}(E)$  is reduced to only one interval. As a consequence,  $\gamma$  can be decomposed in a sequence  $[x, v_0], E_1, E_2, \dots, E_k, [v_k, y]$  where  $[x, v_0]$  and  $[v_k, y]$  are pieces of edges containing  $x$  and  $y$ , respectively, and  $E_1 = [v_0, v_1], E_2 = [v_1, v_2], \dots, E_k = [v_{k-1}, v_k]$  are pairwise distinct edges of  $G$  contained in  $B$ . It follows from Lemma 4 that the lengths of the edges  $E_1, \dots, E_k$  and of  $[x, v_0]$  and  $[v_k, y]$  are upper bounded by  $4\alpha$ . As a consequence, the length of  $\gamma$  is

**Fig. 1** Tightness of the bound in Lemma 3: there are 3 edges of length at most  $4\alpha$  and the diameter of  $B$  is equal to  $20\alpha$ . The range of the distances from  $r$  to the points on the red curve is  $[d - \alpha, d + \alpha]$



**Fig. 2** The path correspondence between two metric spaces  $X$  and  $Y$



upper bounded by  $4(k + 2)\alpha$  which is itself upper bounded by  $4(N_E(4\alpha) + 2)\alpha$  since the edges  $E_1, \dots, E_k$  are pairwise distinct. It follows that  $d_G(x, y) \leq 4(N_E(4\alpha) + 2)\alpha$ .  $\square$

Figure 1 shows that the bound of Lemma 3 is tight.

**Lemma 5** Let  $X$  and  $Y$  be compact geodesic metric spaces and  $C \subset X \times Y$  be an  $\varepsilon'$ -correspondence between them. Assume  $(x_0, y_0) \in C(X, Y)$ , we define functions  $d_{x_0}(\cdot) = d_X(x_0, \cdot)$  and  $d_{y_0}(\cdot) = d_Y(y_0, \cdot)$  in  $X$  and  $Y$ , respectively. Then for any path  $\gamma_x$  in  $X$  connecting  $x_a, x_b \in X$ , we can find a path  $\gamma_y$  in  $Y$  such that its end points  $y_a, y_b$  are corresponding to  $x_a, x_b$ . And furthermore

$$[\min_{y \in \gamma_y} d_{y_0}(y), \max_{y \in \gamma_y} d_{y_0}(y)] \subset [\min_{x \in \gamma_x} d_{x_0}(x) - 2\varepsilon', \max_{x \in \gamma_x} d_{x_0}(x) + 2\varepsilon'].$$

*Proof* Let  $\varepsilon > \varepsilon' > 0$  and  $u, l$  be the maximum and minimum of  $d_{x_0}$  restricted to  $\gamma_x$ . Since  $C$  is an  $\varepsilon'$ -correspondence for any  $x \in \gamma_x$ , there exists a point  $(x, y) \in C$  such that  $d_{x_0}(x) - \varepsilon' \leq d_{y_0}(y) \leq d_{x_0}(x) + \varepsilon'$ . As illustrated in Fig. 2, 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  $x_1 = x_a, x_2, \dots, x_n = x_b$  of points in  $\gamma_x$  such that for any  $i = 1, \dots, n - 1$  we have  $d_X(x_i, x_{i+1}) < \varepsilon - \varepsilon'$ . If  $(x_i, y_i) \in C$ , then we have

$d_Y(y_i, y_{i+1}) < \varepsilon - \varepsilon' + \varepsilon' = \varepsilon$ . As a consequence, since  $l - \varepsilon < l - \varepsilon' < d_{y_0}(y_i) < u + \varepsilon' < u + \varepsilon$ , the shortest geodesic connecting  $y_i$  to  $y_{i+1}$  in  $G$  remains in the set  $d_{y_0}^{-1}([l - 2\varepsilon, u + 2\varepsilon])$  and connecting these geodesics for all  $i = 1, \dots, n - 1$  we get a continuous path from  $y_a$  to  $y_b$  in  $d_r^{-1}([l - 2\varepsilon, u + 2\varepsilon])$ . Now decreasing  $\varepsilon$  to  $\varepsilon'$ , we finish the construction.  $\square$

As a corollary, we have the following theorem.

**Theorem 2** *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  $d_{x_0} = d_X(x_0, \cdot) : X \rightarrow [0, +\infty)$  be the distance function to  $x_0$ . Then for  $d \geq \alpha \geq 0$ , the diameter of any connected component  $L$  of  $d_{x_0}^{-1}([d - \alpha, d + \alpha])$  satisfies*

$$\text{diam}(L) \leq 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon) + \varepsilon,$$

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  $8\varepsilon$  is smaller than the length of the shortest edge of  $G$ , then  $\text{diam}(L) < 17\varepsilon$ .

*Proof* Let  $\varepsilon' > 0$  be such that  $d_{GH}(X, G) < \varepsilon' < \varepsilon$ . Let  $C \subset X \times G$  be an  $\varepsilon'$ -correspondence between  $X$  and  $G$  and  $(x_0, r) \in C$ . We denote by  $d_r = d_G(r, \cdot) : G \rightarrow [0, +\infty)$  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  $x_b$ . Following Lemma 5, 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 3 that  $d_G(y_a, y_b) \leq 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon)$  and since  $C$  is an  $\varepsilon'$ -correspondence (and so an  $\varepsilon$ -correspondence), we get  $d_X(x_a, x_b) < 4(2 + N_E(4(\alpha + 2\varepsilon)))(\alpha + 2\varepsilon) + \varepsilon$ .  $\square$

From Theorems 1 and 2, we obtain the following results for the Reeb and  $\alpha$ -Reeb graphs.

**Theorem 3** *Let  $(X, d_X)$  be a compact connected path metric space, and let  $r \in X$  be a fixed base point such that the metric Reeb graph  $(G, d_G)$  of the function  $d = d_X(r, \cdot) : X \rightarrow \mathbb{R}$  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  $8\varepsilon$ . In particular, if  $X$  is at distance less than  $\varepsilon$  from a metric graph with shortest edge length larger than  $8\varepsilon$ , then  $d_{GH}(X, G) < 17(\beta_1(G) + 1)\varepsilon$ .

**Theorem 4** *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  $[0, \text{Diam}(X)]$  by open intervals of length at most  $\alpha$  such that the  $\alpha$ -Reeb graph  $G_\alpha$  associated to  $\mathcal{I}$  and the function  $d = d_X(r, \cdot) : X \rightarrow \mathbb{R}$  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  $\varepsilon$  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  $\alpha$ -Reeb graph  $G$  of  $(X, d)$  recovers some topology of  $X$ .

**Theorem 5** *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  $\mathcal{I} = \{[0, 2\alpha), (i\alpha, (i+2)\alpha) \mid 1 \leq i \leq m\}$  covers the segment  $[0, \text{Diam}(X)]$  such that the  $2\alpha$ -Reeb graph  $G$  associated to  $\mathcal{I}$  and the function  $d = d_X(r, \cdot) : X \rightarrow \mathbb{R}$  is a finite graph. If no edges of  $G'$  are shorter than  $L$  and no loops (lengths) of  $G'$  are shorter than  $2L$  with  $L \geq 32\alpha + 9\varepsilon$ , then we have  $G$  and  $G'$  are homotopy equivalent.*

Our strategy of proving Theorem 5 is to construct some open covers for  $X$  and  $G'$  and relate the  $\alpha$ -Reeb graph  $G$  and the graph  $G'$  to the nerves of the open covers. Specifically, we construct an initial open cover  $\mathcal{V}_0$  of  $X$  whose nerve  $N(\mathcal{V}_0)$  is homotopy equivalent to  $G$ . Then we obtain a new open cover  $\tilde{\mathcal{V}}$  of  $X$  by merging certain elements in  $\mathcal{V}_0$  while preserving the homotopy type of the nerve of the open cover, i.e.,  $N(\mathcal{V}_0)$  and  $N(\tilde{\mathcal{V}})$  are homotopy equivalent. Based on the open cover  $\tilde{\mathcal{V}}$ , we construct an open cover  $\tilde{\mathcal{U}}$  for  $G'$  whose nerve  $N(\tilde{\mathcal{U}})$  is isomorphic to  $N(\tilde{\mathcal{V}})$  as graphs and at the same time is homotopy equivalent to  $G'$ . In the following, we describe the constructions of the above open covers for  $X$  and  $G'$  and show the above claimed relations between them.

Since  $d_{GH}(X, G') < \varepsilon$ , there exists an  $\varepsilon$ -correspondence between the two spaces, denoted by  $C(X, G')$ . 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  $C(U)$  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  $b : G' \rightarrow \mathbb{R}$  by  $b(g) = d_{G'}(g_r, g)$ . Let  $N = \{g_{n_1}, g_{n_2}, \dots, g_{n_p}\}$  be the vertices of  $G'$ , i.e.,  $N$  is the set of vertices whose degree is not equal to two. From the hypotheses 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  $M = \{g_{m_1}, \dots, g_{m_q}\}$ . Note that 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 split into at most two edges in  $G'$ .

### 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 by  $V_k^l$  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 by  $N(\mathcal{V}_0)$ , is a graph. 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. 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 6** *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 hypotheses of Theorem 5, we have  $j - k \geq 15$ .*

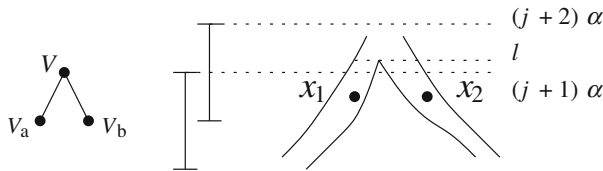
*Proof* First notice that  $j > k$ . Let  $x_1 \in V_k^l \cap d^{-1}(k\alpha, (k+1)\alpha)$  and  $x_2 \in V_j^i \cap d^{-1}((j+1)\alpha, (j+2)\alpha)$ . From the hypotheses, there are two different paths  $\beta_1, \beta_2$  connecting  $x_1$  to  $x_2$  so that  $\beta_1 \cap d^{-1}((k+1)\alpha, (j+1)\alpha)$  and  $\beta_2 \cap d^{-1}((k+1)\alpha, (j+1)\alpha)$  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  $i = 1, 2$ . Following Lemma 5, the path  $\beta_i$  in  $X$  for  $i = 1, 2$  traces out a simple path  $\gamma_i$  in  $G'$  connecting  $g_1$  to  $g_2$  so that  $\gamma_i$  lies in  $b^{-1}(k\alpha - \varepsilon, (j+2)\alpha + \varepsilon)$ . One can verify that  $\gamma_1$  and  $\gamma_2$  are two different paths due to the fact that  $\beta_1$  and  $\beta_2$  pass through different connected components of  $d^{-1}((k+1)\alpha, (j+1)\alpha)$  and thus form a loop in  $G'$ , denoted by  $\gamma$ . We have  $b(\gamma) \subset (k\alpha - \varepsilon, (j+2)\alpha + \varepsilon)$ .

We claim that the range of the function  $b$  restricted to any loop, in particular  $\beta$ , covers an interval with the length at least  $\frac{L}{2}$ . If the claim holds, then we have  $(j-k+2)\alpha + 2\varepsilon \geq \frac{L}{2}$ , which implies  $j - k \geq 15$  from the hypothesis  $L \geq 32\alpha + 9\varepsilon$  of Theorem 5. Indeed, if  $\beta$  contains at least two vertices in  $N$ , then it is obvious that the range of the function  $b$  restricted to  $\gamma$  covers an interval with the length at least  $\frac{L}{2}$  as any original edge of  $G'$  is longer than  $L$ . Now consider the case where  $\gamma$  contains one vertex in  $N$ , say  $g_a$ . If  $\gamma$  does not contain  $g_r$ , then there is exactly one local maximum on  $\gamma$ , say  $g_b$ . If  $\gamma$  contains  $g_r$ , then let  $g_b = g_r$ . The removal of  $g_a$  and  $g_b$  cuts  $\gamma$  into two pieces. Along either piece, the function  $b$  has at most one local maximum. The length of  $\gamma$  is longer than  $2L$ . We have  $b(\gamma)$  that covers an interval with length longer than  $L/2$ . Finally, if  $\beta$  contains no vertex in  $N$ , then  $G'$  is a single loop  $\gamma$  and the claim obviously holds.  $\square$

In the following, we modify this open cover by merging while preserving the homotopy type of its nerve. The main purpose of the merging operation is to relate the open cover of  $X$  to the open cover of  $G'$  as constructed in Sect. 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  $b(g) \geq \frac{\alpha}{2}$ , then there exists a unique positive integer  $k'$  s.t.  $k'\frac{\alpha}{2} \leq b(g) < (k'+1)\frac{\alpha}{2}$ . Let  $k = \lfloor \frac{k'+1}{2} \rfloor - 1 \geq 0$ , and one can verify that  $(k + \frac{1}{2})\alpha \leq b(g) \leq (k + \frac{3}{2})\alpha$ . 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  $d_X(x_i, x_0) \geq |d(x_i) - d(x_0)| \geq \frac{\alpha}{2} - \varepsilon$ . Then  $d_X(x_1, x_2) \geq \alpha - 2\varepsilon$  which contradicts to the fact that  $d_X(x_1, x_2) \leq d_{G'}(g, g) + \varepsilon \leq \varepsilon$ . 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.,





**Fig. 3** Left:  $V$  has two lower neighboring vertices; Right: Illustration of  $V \cup V_a \cup V_b$

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

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 by  $V(g)$  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  $V_k^l$  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 the critical open sets and the regular open sets.

**Lemma 7** *Under the hypotheses of Theorem 5, we have for any vertex  $g \in M \cup N$ ,*

- (i)  $d(V(g)) \subset [s\alpha, (s+4)\alpha]$  for some integer  $s \geq 0$ , and
- (ii) for any point  $x \in \bigcup_{V \in \mathcal{V}_0 \setminus V(g)} V$  and any  $g_x \in C(x) \subset G'$ ,  $d_{G'}(g, g_x) \geq \frac{\alpha}{2} - 2\varepsilon$ .

*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  $x \in \bigcup_{V \in \mathcal{V}_0 \setminus V(g)} V$ , we have  $d(x) > 3\alpha$  and  $b(g_x) > 3\alpha - \varepsilon$ . Thus  $d_{G'}(g, g_x) \geq |b(g_x) - b(g)| > 3\alpha - \varepsilon - \frac{\alpha}{2} > \frac{\alpha}{2} - 2\varepsilon$ . Now consider the case where  $b(g) \geq \frac{\alpha}{2}$ . Assume  $V_k^l$  is the center of  $V(g)$ . If  $d(x) \notin I_k$ , then  $d_X(x, y) \geq \frac{\alpha}{2} - \varepsilon$  for any point  $y \in C(g)$  from the construction of  $V(g)$ , which implies  $d_{G'}(g_x, g) \geq \frac{\alpha}{2} - 2\varepsilon$ . 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  $d_X(x, y) > d_X(x_0, y) \geq \frac{\alpha}{2} - \varepsilon$  and  $d_{G'}(g_x, g) \geq d_X(x, y) - \varepsilon \geq \frac{\alpha}{2} - 2\varepsilon$ . This proves the lemma.  $\square$

**Lemma 8** *For any regular open set  $V \in \mathcal{V}$ ,  $V$  is also an open set in  $\mathcal{V}_0$ . Moreover, under the hypotheses of Theorem 5, it is of degree two in the nerve of  $N(\mathcal{V}_0)$  with one neighboring vertex higher than  $V$  and one neighboring vertex lower than  $V$ .*

*Proof* We prove the lemma by contradiction. Assume  $V \in \mathcal{V}_0 \setminus \bigcup_{g \in M \cup N} V(g)$  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 both  $V_a$  and  $V_b$  have non-empty intersection with  $V$ , there exists a path in  $d^{-1}((j-1)\alpha, (j+2)\alpha)$ . Now let  $l = \inf\{s : \text{there exists a path connecting } x_a, x_b \text{ in } d^{-1}((j-1)\alpha, s] \cap (V \cup V_a \cup V_b)\}$ . As illustrated in Fig. 3,  $V_a, V_b$  are disconnected, thus we have  $(j+1)\alpha \leq l < (j+2)\alpha$ .

We can choose two points  $x_1, x_2 \in V \cup V_a \cup V_b$  from a path connecting  $x_a$  and  $x_b$  such that  $d(x_1) = d(x_2) = l - 2\varepsilon$ , and  $x_1, x_2$  are disconnected in  $d^{-1}([l - 2\varepsilon, l]) \cap (V \cup V_a \cup V_b)$  but are connected by a path, say  $\beta$ , in  $d^{-1}([l - 2\varepsilon, l]) \cap (V \cup V_a \cup V_b)$ . Obviously  $d_X(x_1, x_2) \geq 2(l - (l - 2\varepsilon)) = 4\varepsilon$ . Let  $g_i \in C(x_i)$  for  $i = 1, 2$ . Then  $b(g_i) \in [l - 3\varepsilon, l - \varepsilon]$  for  $i = 1, 2$  and  $d_{G'}(g_1, g_2) \geq d_X(x_1, x_2) - \varepsilon \geq 3\varepsilon$ . Following Lemma 5, the path  $\beta$  traces out a simple path in  $G'$  denoted by  $\gamma$  connecting  $g_1$  and  $g_2$  which lies in  $b^{-1}(l - 4\varepsilon, l + 2\varepsilon)$ . We claim that  $\gamma$  must contain a vertex  $g_c$  in  $M \cup N$ . If not,

$$\begin{aligned} d_{G'}(g_1, g_2) &= |b(g_1) - b(g_2)| \\ &\leq |b(g_1) - d(x_1) + d(x_2) - b(g_2)| \\ &\leq |(b(g_1) - b(g_r)) - (d(x_1) - d(r))| \\ &\quad + |(d(x_2) - d(r)) - (b(g_2) - b(g_r))| \\ &= |d_{G'}(g_1, g_r) - d_X(x_1, r)| + |d_{G'}(g_2, g_r) - d_X(x_2, r)| \\ &\leq 2\varepsilon. \end{aligned}$$

This contradicts to the fact that  $d_{G'}(g_1, g_2) \geq 3\varepsilon$ . From the construction of  $\gamma$ , there exists a point  $g \in \gamma$  so that  $d_{G'}(g, g_c) \leq \varepsilon$  and there exists a point  $x \in \beta \cap C(g)$ . Then for any point  $x_c \in C(g_c)$ , we have  $d_X(x, x_c) \leq 2\varepsilon$ . For any vertex  $x_0 \in V \cap d^{-1}(l)$ , since  $x_0$  and  $x$  are connected in  $d^{-1}([l - 2\varepsilon, l]) \cap (V \cup V_a \cup V_b)$ ,  $d_X(x_0, x) \leq 25\varepsilon$  from Theorem 2, and therefore  $d_X(x_0, x_c) \leq 27\varepsilon$ . However, since  $x_0 \in V$  which is regular,  $x \notin V(g_c)$ . From Lemma 7, we have  $d_X(x_0, x_c) \geq \frac{\alpha}{2} - 2\varepsilon > 27\varepsilon$ . This is a contradiction. Therefore,  $V$  cannot have more than one neighboring vertices that are lower than  $V$ .

Using a similar argument, we can also prove that  $V$  cannot have more than one neighboring vertices that are higher than  $V$ .  $\square$

We now perform a second step of merging. Two critical open sets  $V(g_1)$  and  $V(g_2)$  in  $N(\mathcal{V})$  are said to be close if there is a simple path  $\gamma$  in the nerve  $N(\mathcal{V}_0)$  connecting the center  $V_{k_1}^{l_1}$  of  $V(g_1)$  and the center  $V_{k_2}^{l_2}$  of  $V(g_2)$  so that  $\gamma$  consists of at most 4 edges. If there is a regular open set along the above path, we say this regular open set connects the critical open sets  $V(g_1)$  and  $V(g_2)$ . We have the following properties for two close critical open sets.

**Lemma 9** *Under the hypotheses of Theorem 5, we have*

- (i) *for any two vertices  $g_{n_1}, g_{n_2} \in N$ ,  $V(g_{n_1})$  and  $V(g_{n_2})$  cannot be close;*
- (ii) *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;*

- (iii) 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 at least one of  $V(g_{m_1})$  and  $V(g_{m_2})$  is close to  $V(g_n)$ . Moreover, there is a path in  $N(\mathcal{V}_0)$  of at most 5 edges connecting the center of  $V(g_n)$  to the center of  $V(g_{m_i})$  for any  $i = 1, 2$ .

*Proof* Let  $V_j^p, V_k^q$  be the centers of  $V(g_1)$  and  $V(g_2)$ , respectively, for  $g_1, g_2 \in N \cup M$ . If  $V(g_1)$  and  $V(g_2)$  are close, then  $|k - j| \leq 4$ . Assume  $j \leq k$ . Then for any  $x_1 \in C(g_1)$  and  $x_2 \in C(g_2)$ , there is a path in  $d^{-1}((j\alpha, (k+2)\alpha))$  connecting  $x_1$  and  $x_2$ . Note that  $k + 2 - j \leq 6$ . We claim that  $V(g_1)$  and  $V(g_2)$  are not close provided that  $d_{G'}(g_1, g_2) > 12\alpha + 9\epsilon$ . Indeed, since  $d_{G'}(g_1, g_2) > 12\alpha + 9\epsilon$ , the range of the function  $b$  restricted to any path connecting  $g_1$  and  $g_2$  in  $G'$  covers an interval of the length at least  $6\alpha + 4.5\epsilon$ . This implies that the range of the function  $d$  restricted to any path in  $X$  connecting  $x_1$  and  $x_2$  covers an interval of the length at least  $6\alpha + 0.5\epsilon$ . This means that  $V(g_1)$  and  $V(g_2)$  cannot be close. Since  $d_{G'}(g_{n_1}, g_{n_2}) \geq L > 12\alpha + 9\alpha$ ,  $V(g_{n_1})$  and  $V(g_{n_2})$  are not close. This proves (i).

Assume  $V(g_m)$  is close to both  $V(g_{n_1})$  and  $V(g_{n_2})$  with  $g_{n_1}, g_{n_2} \in N$ . We have  $d_{G'}(g_{n_1}, g_{n_2}) \leq d_{G'}(g_m, g_{n_1}) + d_{G'}(g_m, g_{n_2}) \leq 24\alpha + 18\epsilon < L$ , which means  $g_{n_1} = g_{n_2}$ . This proves (ii).

We now prove (iii). Since at most one vertex in  $M$  is added into an original edge of  $G'$ , any path in  $G'$  connecting  $g_{m_1}$  and  $g_{m_2}$  passes through at least one vertex from  $N$ . Furthermore, let  $\gamma$  be a geodesic in  $G'$  connecting  $g_{m_1}$  and  $g_{m_2}$ . If  $\gamma$  passes more than one vertices in  $N$ ,  $d_{G'}(g_{m_1}, g_{m_2}) \geq L > 12\alpha + 9\epsilon$ , which contradicts to the fact that  $V(g_{m_1})$  and  $V(g_{m_2})$  are close. Therefore  $\gamma$  contains exactly one vertex in  $N$ . Denote this vertex by  $g_n$ . Let  $V_{k_1}^{l_1}$  and  $V_{k_2}^{l_2}$  be the centers of  $V(g_{m_1})$  and  $V(g_{m_2})$ , respectively, and  $\delta$  be the simple path from  $V_{k_1}^{l_1}$  to  $V_{k_2}^{l_2}$  in  $N(\mathcal{V}_0)$  so that  $\delta$  consists of at most 4 edges, or equivalently at most five elements in  $\mathcal{V}_0$ .

Recall that  $V(g_n)$  consists of a subset of  $\mathcal{V}_0$ . We claim that  $\delta$  must pass through an element in  $V(g_n)$ .

If the claim holds, it is easy to verify that at least one of  $V(g_{m_1})$  and  $V(g_{m_2})$  is close to  $V(g_n)$ . In addition, if we let  $V_k^l$  be the center of  $V(g_n)$ , then there is a path in  $N(\mathcal{V}_0)$  with at most 5 edges connecting  $V_{k_i}^{l_i}$  and  $V_k^l$  for any  $i = 1, 2$ . This proves (iii).

It remains to show the above claim. We prove by contradiction. If we let

$$V(\delta) = \{V \in \mathcal{V}_0 : V \text{ is on the path of } \delta\},$$

then  $V(g_n)$  as a subset of  $\mathcal{V}_0$  does not intersect with  $V(\delta)$ . We have  $C(g_{m_1})$  and  $C(g_{m_2})$  being contained in  $V_{k_1}^{l_1}$  and  $V_{k_2}^{l_2}$  respectively. For any  $x_1 \in C(g_{m_1})$  and any  $x_2 \in C(g_{m_2})$ , there is a path  $\beta$  in  $X$  connecting  $x_1$  and  $x_2$  so that  $\beta$  is contained in  $\bigcup_{V \in V(\delta)} V$ . From Lemma 7, for any  $x \in \beta$  and any  $g_x \in C(x)$ ,  $d_{G'}(g_x, g_n) \geq \frac{\alpha}{2} - 2\epsilon$ . From the construction in the proof of Theorem 2, the path  $\beta$  can trace out a simple path  $\gamma'$  in  $G'$  connecting  $g_{m_1}$  and  $g_{m_2}$  so that for any point  $g \in \gamma'$ ,  $d_{G'}(g, g_n) \geq \frac{\alpha}{2} - 3\epsilon$ . This means that  $\gamma$  and  $\gamma'$  form a loop in  $G'$ . Since  $\beta$  is contained in  $\bigcup_{V \in V(\delta)} V$ ,  $d(\beta)$  is contained in an interval with the length of  $6\alpha$ , which implies that  $b(\gamma')$  is contained in an interval with the length of  $6\alpha + 4\epsilon$ . Thus the length of  $\gamma'$  is at most  $2(6\alpha + 4\epsilon)$  as it passes through at most one vertex in  $N$ . Since  $\gamma$  is a geodesic in  $G'$  connecting

$g_{m_1}$  and  $g_{m_2}$ , the length of the above loop is at most  $4(6\alpha + 4\varepsilon)$ , which contradicts to the hypotheses of Theorem 5. This proves the above claim.  $\square$

We say  $g_a, g_b \in M \cup N$  are equivalent, denoted by  $g_a \sim_c g_b$ , if there exists a finite sequence  $g_a = g_1, g_2, \dots, g_k = g_b$  such that  $V(g_i)$  and  $V(g_{i+1})$  are close for any  $i = 1, \dots, k-1$ . This is an equivalence relation. From Lemma 9iii, if an equivalence class contains at least two vertices in  $M \cup N$ , it must contain a vertex in  $N$ . We have the following lemma:

**Lemma 10** *Under the hypotheses of Theorem 5, an equivalence class contains at most one vertex from  $N$ .*

*Proof* If not, assume  $g_{n_1} \neq g_{n_2}$  and  $g_{n_1} \sim_c g_{n_2}$ . Let  $g_{n_1} = g_1, g_2, \dots, g_k = g_{n_2}$  be a sequence so that  $V(g_i)$  and  $V(g_{i+1})$  are close for any  $i = 1, \dots, k-1$ . Without loss of generality, we can further assume  $g_i \in M$  for  $i = 2, \dots, k-1$ .

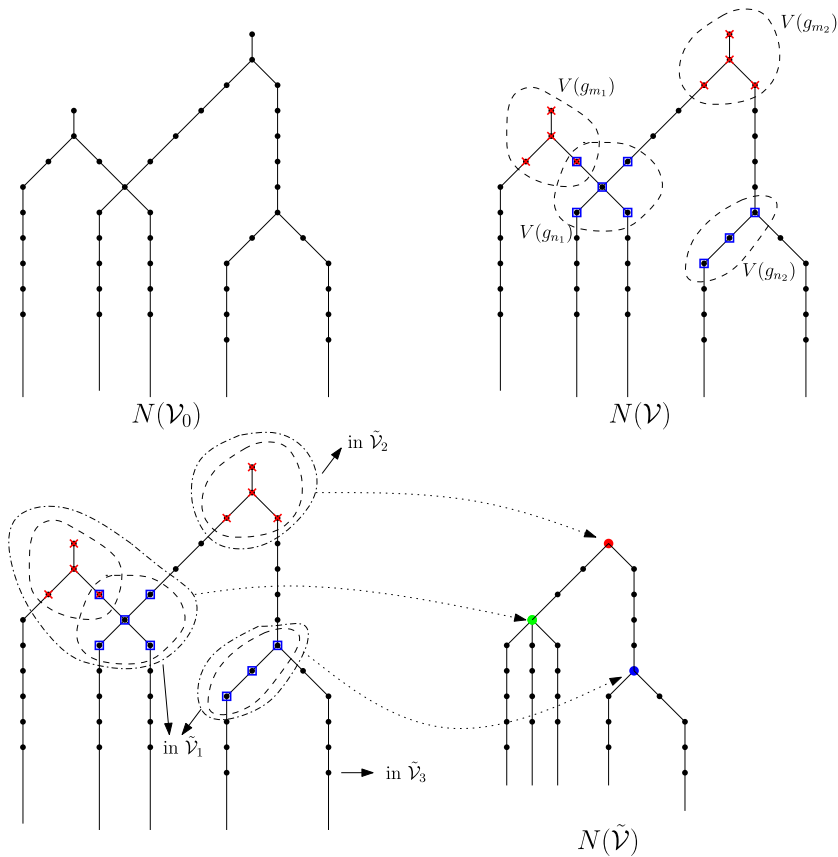
We first show that  $k > 5$ . Assume not. Let  $V_{k_i}^{l_i}$  be the center of  $V(g_i)$  for  $i = 1, \dots, k$ . From Lemma 9iii, there is a path in  $N(\mathcal{V}_0)$  with at most  $2 \times 5 = 10$  edges connecting  $V_{k_1}^{l_1}$  to  $V_{k_5}^{l_5}$ . Thus for any  $x_1 \in C(g_{n_1})$  and any  $x_2 \in C(g_{n_2})$ , there is a path  $\beta$  in  $X$  connecting  $x_1$  and  $x_2$  so that  $d(\beta)$  is contained in an interval with the length at most  $12\alpha$ . The path  $\beta$  traces out a path  $\gamma$  in  $G'$  connecting  $g_{n_1}$  and  $g_{n_2}$  so that  $b(\gamma)$  is contained in an interval with the length at most  $12\alpha + 4\varepsilon$ , which implies  $d_{G'}(g_{n_1}, g_{n_2}) \leq 2(12\alpha + 4\varepsilon)$ . This contradicts to the hypothesis concerning the lengths of the edges in  $G'$ .

Now we assume  $k > 5$ . Since  $V(g_3)$  and  $V(g_4)$  are close and  $g_3, g_4 \in M$ , from Lemma 9iii, there exists a  $g_n \in N$  so that  $V(g_n)$  is close to at least one of  $V(g_3)$  and  $V(g_4)$ . Assume  $V(g_n)$  is close to  $V(g_3)$ . If  $V(g_n) \neq V(g_{n_1})$ , we obtain a sequence of  $g'_1 = g_{n_1}, g'_2 = g_2, g'_3 = g_3, g'_4 = g_n$  so that  $V(g'_i)$  and  $V(g'_{i+1})$  are close for any  $i = 1, \dots, 3$ . If  $V(g_n) \neq V(g_{n_2})$ , we obtain a sequence  $g'_1 = g_n, g'_2 = g_3, \dots, g'_{k-1} = g_{n_2}$  so that  $V(g'_i)$  and  $V(g'_{i+1})$  are close for any  $i = 1, \dots, k-2$ . In either case, the new sequence has a length less than  $k$ . Similarly, we can obtain a shorter sequence if  $V(g_n)$  is close to  $V(g_4)$ . One can keep shortening the sequence so that its length is no longer than 5, which however has been proven to be impossible. This proves the lemma.  $\square$

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  $V(g) \subset \mathcal{V}$  with  $g \sim_c g_n$ , and (3) any regular open set  $V \subset \mathcal{V}$  connecting two critical open sets which are equivalent to  $g_n$ . We abuse the notation and also denote by  $\tilde{V}(g_n)$  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)  $\tilde{\mathcal{V}}_2 = \{V(g) : g \in M \text{ and } V(g) \not\subset \tilde{\mathcal{V}}_N\}$ ;
- (3)  $\tilde{\mathcal{V}}_3 = \{V \in \mathcal{V} : V \text{ is regular and } V \not\subset \tilde{\mathcal{V}}_N\}$ .

Figure 4 shows different types of elements in  $\tilde{\mathcal{V}}$ . We summarize the properties for the open cover  $\tilde{\mathcal{V}}$  in the following corollary, which follows from Lemmas 7–10.



**Fig. 4** Illustration of the merging strategy: From *top-left* to *top-right* (First step): We select the unions of open sets from  $\mathcal{V}_0$  based on the critical points w.r.t.  $b(g)$  in  $G'$  and merge them, respectively. From *top-right* to *bottom-left* (Second step): We further merge the unions which are mutually close. *Bottom-right* the nerve of the merging result  $N(\tilde{\mathcal{V}})$

**Corollary 1** *Under the hypotheses of Theorem 5, 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  $\tilde{V}_1, \tilde{V}_2 \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$ , any path in the 1-skeleton of the nerve  $N(\tilde{\mathcal{V}})$  connecting  $\tilde{V}_1, \tilde{V}_2$  consists of at least two elements from  $\tilde{\mathcal{V}}_3$ .
- Any open set  $\tilde{V} \in \tilde{\mathcal{V}}_3$  is also a regular element in  $\mathcal{V}$  and thus an element in  $\mathcal{V}_0$ . Moreover, any point  $g \in C(\tilde{V}) \subset G'$  is at least  $\frac{\alpha}{2} - 2\varepsilon$  away from any vertex of  $G'$ .

**Proposition 4** *Under the hypotheses of Theorem 5,  $N(\tilde{\mathcal{V}})$  and  $N(\mathcal{V}_0)$  are homotopy equivalent.*

*Proof* We obtain the elements in the open covering  $\tilde{\mathcal{V}}$  by merging a subset of open sets in  $\mathcal{V}_0$ . Think of any element  $\tilde{V} \in \tilde{\mathcal{V}}$  as a subset of  $\mathcal{V}_0$ . The nerve  $N(\mathcal{V}_0)$  restricted to  $\tilde{V}$

is a subgraph of  $N(\mathcal{V}_0)$ , whose vertex set is  $\tilde{V}$  and edge set includes the edges in  $N(\mathcal{V}_0)$  with both endpoints in  $\tilde{V}$ . We call this subgraph the nerve of  $\tilde{V}$ , denoted by  $N(\tilde{V})$ . The nerve of  $N(\tilde{V})$  as a topological space is the quotient space  $N(\mathcal{V}_0)/\bigcup_{\tilde{V} \in \tilde{\mathcal{V}}} N(\tilde{V})$ . From [30, Proposition 0.17], it is sufficient to show that  $N(\tilde{V})$  is a tree for any  $\tilde{V} \in \tilde{\mathcal{V}}$ .

For  $\tilde{V} \in \tilde{\mathcal{V}}_2 \cup \tilde{\mathcal{V}}_3$ ,  $N(\tilde{V})$  is obviously a tree. Consider  $\tilde{V} \in \tilde{\mathcal{V}}_1$ . There exists a  $g_n \in N$  so that  $V(g_n) \subset \tilde{V}(g_n) \in \tilde{V}$ . Let  $V_s^I$  be the center of  $V(g_n)$ . For any  $g_m \in M$  and  $V(g_m) \subset \tilde{V}(g_n)$ , if  $v_{s'}^I$  be the center of  $V(g_m)$ , then from Lemma 9,  $|s - s'| < 5$ . Therefore, if  $V_l^I$  is the element in  $\tilde{V}$  with the smallest sub-index and  $V_h^J$  is the element in  $\tilde{V}$  with the largest sub-index, then we have  $|h - l| \leq 5 + 5 + 2 + 2 = 14$ . From Lemma 6, there is no loop in the subgraph  $N(\tilde{V})$ . This proves the proposition.  $\square$

## 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 an open set  $V \in \mathcal{V}_0$ , we construct a connected open set  $U_V \subset G'$  so that  $C(V) \subset U_V$  as follows: Let  $l = \min\{d(V)\}$  and  $u = \max\{d(V)\}$ . We have  $u - l \leq 2\alpha$ . Let  $\tilde{U} = b^{-1}([l - 2\varepsilon, u + 2\varepsilon])$ , and then  $C(V) \subset \tilde{U}$ . Since  $u - l + 4\varepsilon < 2\alpha + 4\varepsilon < \frac{L}{4}$ , one can verify that there is no loop in  $\tilde{U}$  and thus  $\tilde{U}$  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  $\beta$  be a path in  $V$  connecting  $x_1$  and  $x_2$ . Following Lemma 5,  $\beta$  can trace out a path  $\gamma$  in  $\tilde{U}$  connecting  $g_1$  and  $g_2$ , which implies that  $C(V)$  is contained in a tree in  $\tilde{U}$ . 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 in  $\mathcal{V}_0$ . 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  $U_{\tilde{V}}$  as the open set of the union of the 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 1, any point in  $C(\tilde{V})$  is at least  $\frac{\alpha}{2} - 2\varepsilon$  away from any vertex in  $M \cup N$  and any point in  $U_{\tilde{V}}$  is at least  $\frac{\alpha}{2} - 4\varepsilon$  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  $\tilde{U}_{\tilde{V}} = U_{\tilde{V}} \cap b^{-1}(p\alpha + 2\varepsilon, (p+2)\alpha - 2\varepsilon)$ , which is also a segment in  $G'$ . 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** *Under the hypotheses of Theorem 5, the nerve  $N(\tilde{\mathcal{V}})$  and the nerve  $N(\tilde{\mathcal{U}})$  are isomorphic as graphs.*

*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 8, any regular set has two neighbors in the nerve  $N(\mathcal{V})$  one lower and one higher. Without loss of generality, assume  $\tilde{V}_i$  is higher than  $\tilde{V}_j$ . We have  $\inf\{d(x) : x \in \tilde{V}_i\} \geq \alpha + \sup\{d(x) : x \in \tilde{V}_j\}$ , which implies  $\inf\{b(g) : g \in \tilde{U}_{\tilde{V}_i}\} \geq \alpha + \sup\{b(g) : g \in \tilde{U}_{\tilde{V}_j}\} - 2\varepsilon > \sup\{b(g) : g \in \tilde{U}_{\tilde{V}_j}\}$ . 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  $d(\tilde{V}_j) = (q\alpha, (q+2)\alpha)$ . If  $|p - q| \geq 2$ , it is obvious that  $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$ . Now assume that  $q - p \leq 1$ , which forces the shortest path connecting  $\tilde{V}_i$  and  $\tilde{V}_j$  in  $N(\tilde{\mathcal{V}})$  to pass through some open set  $\tilde{V} \in \tilde{\mathcal{V}}_1 \cup \tilde{\mathcal{V}}_2$ . By Lemma 7, for any  $g_i \in C(\tilde{V}_i)$   $d_{G'}(g_i, g) \geq \frac{\alpha}{2} - 2\varepsilon$  and for any  $g_j \in C(\tilde{V}_j)$   $d_{G'}(g_j, g) \geq \frac{\alpha}{2} - 2\varepsilon$  for any vertex  $g \in M \cup N$  such that  $V(g) \in \tilde{V}$ . Thus  $d_{G'}(g_i, g_j) \geq \alpha - 4\varepsilon$ , 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  $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$ .

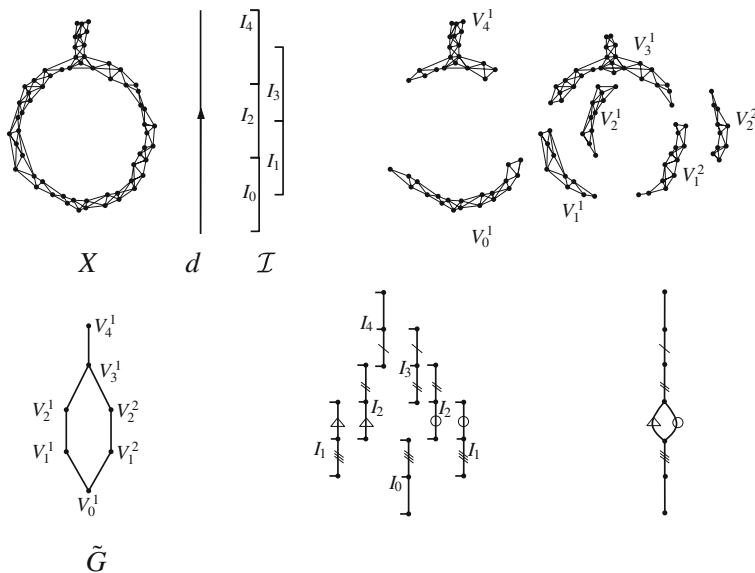
First assume that  $\tilde{V}_i$  and  $\tilde{V}_j$  have a non-empty intersection. As  $\tilde{V}_j \in \tilde{\mathcal{V}}_3$ , it is a regular open set in  $\mathcal{V}$  which has one higher neighboring vertex and one lower neighboring vertex in  $N(\mathcal{V}_0)$ . Since  $\tilde{V}_j$  is regular, we have  $d(\tilde{V}_j) = (p\alpha, (p+2)\alpha)$  for some integer  $p > 0$ . We know that  $\tilde{V}_i$  consists of a subset of open sets in  $\mathcal{V}_0$  and let  $V \in \tilde{V}_i$  be the open set in  $\mathcal{V}_0$  so that  $V \cap \tilde{V}_j \neq \emptyset$ . Without loss of generality, assume  $V$  is the higher neighboring vertex of  $\tilde{V}_j$  and we have  $d(V \cap \tilde{V}_j) \supset ((p+1)\alpha, (p+2)\alpha)$ . We choose a point in  $x \in \tilde{V}_j \cap V$  so that  $d(x) = (p+2)\alpha - 4\varepsilon$ . Since  $b(C(x)) \subset ((p+2)\alpha - 5\varepsilon, (p+2)\alpha - 3\varepsilon)$ ,  $C(x) \in \tilde{U}_{\tilde{V}_j} \cap \tilde{U}_{\tilde{V}_i}$  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  $\beta$  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 scenario is that  $\beta$  contains no intermediate open sets. In this worst scenario, due to the shrinking operation on  $\tilde{U}_{\tilde{V}_j}$ , one can verify that  $\tilde{U}_{\tilde{V}_i} \cap \tilde{U}_{\tilde{V}_j} = \emptyset$ .  $\square$

**Proposition 6** Under the hypotheses of Theorem 5,  $N(\tilde{\mathcal{U}})$  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





**Fig. 5** Illustration of the different steps of the algorithm for computing  $\alpha$ -Reeb graph. In the disjoint union of copies of intervals, the subintervals marked with same labels are identified in the  $\alpha$ -Reeb graph

and thus is a tree and in particular is contractible. Furthermore, the union of any two elements of  $\tilde{\mathcal{U}}$  does not contain 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 that  $N(\tilde{\mathcal{U}})$  is homotopy equivalent to  $G'$ .  $\square$

**Proof of Theorem 5** From Propositions 4–6, it remains to show that the nerve  $N(\mathcal{V}_0)$  is homotopy equivalent to the  $\alpha$ -Reeb graph  $G$ . Indeed, we represent each node  $V_k^l$  in  $N(\mathcal{V}_0)$  using a copy of the interval  $I_k$ . If  $V_{k_1}^{l_1}$  and  $V_{k_2}^{l_2}$  with  $k_1 < k_2$  are the endpoints of an edge in  $N(\mathcal{V}_0)$ , then we glue the upper half of  $I_{k_1}$  to the lower half of  $I_{k_2}$ . We identify any two points which are glued together directly or indirectly. By definition, the  $\alpha$ -Reeb graph is the quotient space of the disjoint union of these intervals—see Fig. 5. From Lemma 6, there is more than one node between the top node and the bottom node of any loop in  $N(\mathcal{V}_0)$ . Thus, we have a one-to-one correspondence between the loops in  $N(\mathcal{V}_0)$  and the loops in the  $\alpha$ -Reeb graph. This proves the theorem.  $\square$

## 6 Algorithm

In this section, we describe an algorithm for computing the  $\alpha$ -Reeb graph for some  $\alpha > 0$ . We assume that the input of the algorithm includes a neighboring graph  $X = (V, E)$ , a function  $l : E \rightarrow \mathbb{R}^+$  specifying the edge length and a parameter  $\alpha$ . 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  $X$  as a Rips graph of the input points with the parameter chosen as a fraction of  $\alpha$ . We assume  $X$  is connected as one can apply the algorithm to each connected component otherwise.

Our algorithm, whose different steps are illustrated in Fig. 5, can be described as follows: In the first step, we fix a node of  $X$  as the root  $r$  and then obtain the distance function  $d : V \rightarrow \mathbb{R}^+$  by computing  $d(v)$  as the graph distance from the node  $v$  to  $r$ . In the second step, we apply the Mapper algorithm [34] to the nodes  $V$  with filter  $d$  to construct a graph  $\tilde{G}$ . Specifically, let  $\mathcal{I} = \{(i\alpha, (i+1)\alpha), ((i+0.5)\alpha, (i+1.5)\alpha) \mid 0 \leq i \leq m\}$  so that  $\bigcup_{I_k \in \mathcal{I}} I_k$  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  $V_k = d^{-1}(I_k)$  and  $V_k^l$  be the  $l$ th component of  $V_k$ . Then  $\{V_k^l\}_{k,l}$  is a cover of  $H$  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  $\{V_k^l\}_{k,l}$ . 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 fact, when we check if  $V_{k_1}^{l_1} \cap V_{k_2}^{l_2} \neq \emptyset$ , we only need to check if their vertices are overlapped or not as we assume the lengths of the edges in  $H$  are fractions of  $\alpha$ .

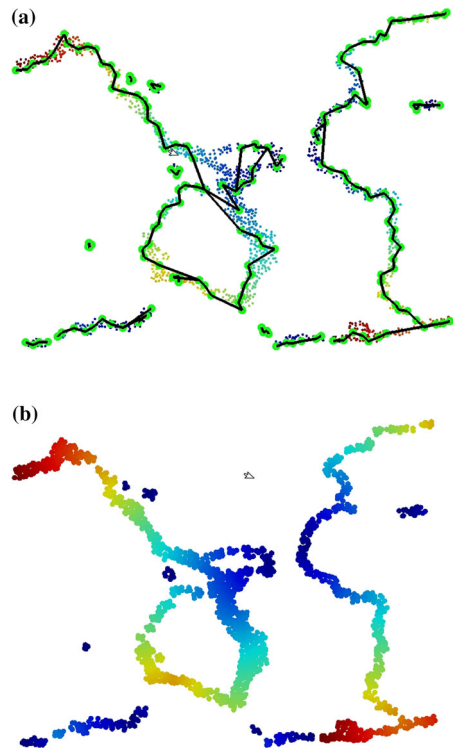
In the final step, we represent each node  $V_k^l$  in  $\tilde{G}$  using a copy of the interval  $I_k$ . As mentioned in the Sect. 3,  $\alpha$ -Reeb graph is a quotient space of the disjoint union of those copies of intervals. Specifically, for an edge in  $\tilde{G}$ , let  $V_{k_1}^{l_1}$  and  $V_{k_2}^{l_2}$  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  $\alpha$ -Reeb graph. Algorithmically, the identification is performed as follows: We split each copy of interval  $I_k$  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  $I_{k_2}$  cannot 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.

## 7 Experiments

In this section, we illustrate the performances of our algorithm on three different datasets. The first dataset was obtained from USGS Earthquake Search [20]. It consists of earthquake epicenter locations collected between 01/01/1970 and 01/01/2010 in the rectangular area between latitudes  $-75^\circ$  and  $75^\circ$  and longitude  $-170^\circ$  and  $10^\circ$  with magnitude greater than 5.0. This raw earthquake dataset contains the coordinates of the epicenters of 12790 earthquakes that are mainly located around geological faults. We follow the procedure described in [1] 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 Fig. 6a. We also set  $\alpha = 4$  and apply

**Fig. 6** Earthquake data  
**a** The distance functions  $d$  on each connected component. The value increases from cold to warm colors. **b** The reconstructed  $\alpha$ -Reeb graph



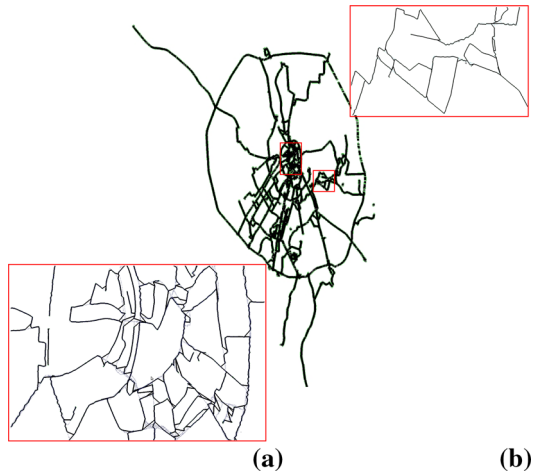
our algorithm to the above data to obtain the  $\alpha$ -Reeb graph. In general, the  $\alpha$ -Reeb graph is an abstract metric graph. In this example, for the purpose of visualization, we use the coordinates of the landmarks to embed the graph into the plane as follows: Recall that for a copy of interval  $I_k$  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 6b shows the embedding of the  $\alpha$ -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  $\alpha$ -Reeb graph.

The second dataset is that of 500 GPS traces tagged “Moscow” from OpenStreetMap [32]. 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  $\varepsilon$ -net on the raw GPS locations with  $\varepsilon = 0.0001$  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 Fig. 7a. Set  $\alpha$  also equal to 0.0004 and compute the  $\alpha$ -Reeb graph. Again, we use the same method as above to embed the  $\alpha$ -Reeb graph into the plane, as shown in Fig. 7b.

To evaluate the quality of our  $\alpha$ -Reeb graph for each dataset, we computed both original pairwise distances, and pairwise distances approximated from the constructed

**Fig. 7** GPS data

**a** The distance functions  $d$  on each connected component. The value increases from cold to warm colors. **b** The reconstructed  $\alpha$ -Reeb graph

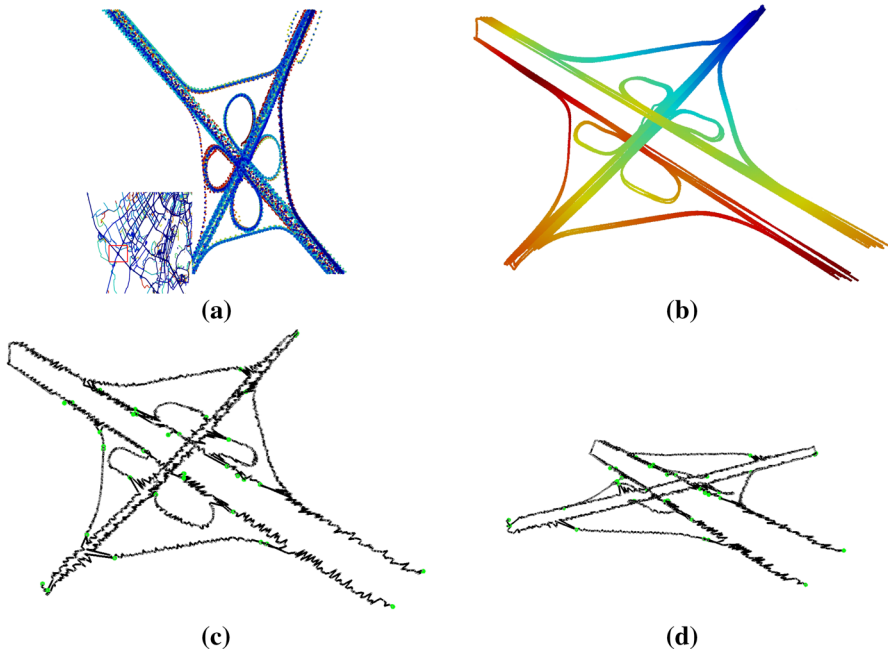
**Table 1** #OP (#OE, #N, #E) stands for the number of original points (original edges, nodes, edges in  $\alpha$ -Reeb graph)

	#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

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

$\alpha$ -Reeb graph. For GPS traces, we randomly select 100 points as the dataset is too big to compute all pairwise distances. We also evaluated the use of  $\alpha$ -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.

The third dataset 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 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 dimensional 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 Fig. 8a. Since GPS records the position based on time, we resample the traces so



**Fig. 8** **a** GPS traces passing through a highway crossing in Moscow. **b** The distance function. **c** and **d** The reconstructed  $\alpha$ -Reeb graph viewed from two perspectives

that the distance between any two consecutive samples is the same among all traces. Then we apply the above algorithm to the resampled traces. Figure 8c and d shows 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 datasets. Here we point out a few possible directions for future work. First, notice that the  $\alpha$ -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 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 Gunnar 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 TopData (ANR-13-BS01-0008); The National Basic Research Program of China (973 Program 2012CB825501); The NSF of China (11271011).

## References

1. Aanjaneya, M., Chazal, F., Chen, D., Glisse, M., Guibas, L., Morozov, D.: Metric graph reconstruction from noisy data. *Int. J. Comput. Geom. Appl.* **22**(04), 305–325 (2012)
2. Abraham, I., Balakrishnan, M., Kuhn F., Malkhi, D., Ramasubramanian, V., Talwar, K.: Reconstructing approximate tree metrics. In: *Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing, PODC*, pp. 43–52. ACM Press, New York, NY (2007)
3. Amenta, N., Bern, M., Eppstein, D.: The crust and the  $\beta$ -skeleton: combinatorial curve reconstruction. *Graph. Models Image Process.* **60**(2), 125–135 (1998)
4. Arias-Castro, E., Donoho, D., Huo, X.: Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Stat.* **34**(1), 326–349 (2006)
5. Bădoiu, M., Indyk, P., Sidiropoulos, A.: Approximation algorithms for embedding general metrics into trees. In: *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms, SODA '07*, pp. 512–521 (2007)
6. Bauer, U., Ge, X., Wang, Y.: Measuring distance between Reeb graphs. [arXiv:1307.2839](https://arxiv.org/abs/1307.2839) (2013)
7. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)
8. Burago, D., Burago, Y.: *A Course in Metric Geometry*. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI (2001)
9. Carlsson, G.: Topology and data. *AMS Bull.* **46**, 255–308 (2009)
10. Chazal, F., Sun, J.: Gromov–Hausdorff approximation of filament structure using Reeb-type graph. In: *Proceedings of the Symposium on Computational Geometry*, pp. 491–500. ACM Press, New York, NY (2014)
11. Chazal, F., de Silva, V., Glisse, M., Oudot, S.: The structure and stability of persistence modules. [arXiv:1207.3674](https://arxiv.org/abs/1207.3674) (2012)
12. Chen, D., Guibas, L., Hershberger, J., Sun, J.: Road network reconstruction for organizing paths. In: *Proceedings 21st ACM-SIAM Symposium on Discrete Algorithms*, pp. 1309–1320. SIAM, Philadelphia (2010)
13. Chepoi, V., Dragan, F., Estellon, B., Habib, M., Vaxès, Y.: Notes on diameters, centers, and approximating trees of delta-hyperbolic geodesic spaces and graphs. *Electron. Notes Discrete Math.* **31**, 231–234 (2008)
14. Choi, E., Bond, N.A., Strauss, M.A., Coil, A.L., Davis, M., Willmer, C.N.A.: Tracing the filamentary structure of the galaxy distribution at  $z \sim 0.8$ . *Monthly Notices Royal Astron. Soc.*, **406**(1), 320–328. [arXiv:1003.3239](https://arxiv.org/abs/1003.3239) (2010)
15. Dey, T., Wenger, R.: Reconstructing curves with sharp corners. *Comput. Geom. Theory Appl.* **19**, 89–99 (2001)
16. Dey, T., Mehlhorn, K., Ramos, E.: Curve reconstruction: connecting dots with good reason. In: *Proceedings of the 15th Symposium on Computational Geometry*, pp. 197–206 (1999)
17. Dey, T.K., Fan, F., Wang, Y.: Graph induced complex on point data. In: *Proceedings of the 29th Annual ACM Symposium on Computational Geometry*, pp. 107–116. ACM Press, New York, NY (2013)
18. Dhamdhere, K., Gupta, A., Räcke, H.: Improved embeddings of graph metrics into random trees. In: *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithm, SODA '06*, pp. 61–69 (2006)
19. Dieudonné, J.: *Foundations of Modern analysis*. Pure and Applied Mathematics. Academic Press, New York (1969)
20. Earthquake search. <http://earthquake.usgs.gov/earthquakes/eqarchives/epic/>. Accessed Sep 2008
21. Edelsbrunner, H., Harer, J.: *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI (2010)
22. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological Persistence and Simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002)
23. Fakcharoenphol, J., Rao, S., Talwar, K.: A tight bound on approximating arbitrary metrics by tree metrics. In: *Proceedings of the 35th ACM Symposium on Theory of Computing, STOC '03*, pp. 448–455. ACM, New York, NY (2003)
24. Ge, X., Safa, I., Belkin, M., Wang, Y.: Data skeletonization via Reeb graphs. *Neural Inf. Process. Syst.* **24**, 837–845 (2011)
25. Genovese, C., Perone-Pacífico, M., Verdinelli, I., Wasserman, L.: On the path density of a gradient field. *Ann. Stat.* **37**(6A), 3236–3271 (2009)

26. Genovese, C., Perone-Pacífico, M., Verdinelli, I., Wasserman, L.: Nonparametric ridge estimation. [arXiv:1212.5156](#) (2012)
27. Genovese, C.R., Perone-Pacífico, M., Verdinelli, I., Wasserman, L.: The geometry of nonparametric filament estimation. *J. Am. Stat. Assoc.*, **107**(498), 788–799 (2012)
28. Gromov, M.: *Metric Structures for Riemannian and Non-Riemannian Spaces*. Modern Birkhäuser Classics. Birkhäuser, Berlin (2002)
29. Harvey, W., Wang, Y., Wenger, R.: A randomized  $O(m \log m)$  time algorithm for computing Reeb graph of arbitrary simplicial complexes. In: *Proceedings of the 26th Annual ACM Symposium on Computational Geometry* (2010). ACM, New York, NY
30. Hatcher, A.: *Algebraic Topology*. Cambridge University Press, New York (2002)
31. Lafon, S.: *Diffusion Maps and Geodesic Harmonics*, PhD. Thesis, Yale University (2004). <https://sites.google.com/site/stefansresearchpapers/>
32. Openstreetmap. <http://www.openstreetmap.org/>. Accessed Jun 2010
33. Parsa, S.: A deterministic  $O(m \log m)$  time algorithm for the Reeb graph. In: *Proceedings of the 2012 Symposium on Computational Geometry*, pp. 269–276 (2012)
34. Singh, G., Mémoli, F., Carlsson, G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: *Proceedings of the Eurographics Symposium on Point-Based Graphics*, pp. 91–100 (2007)
35. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000)
36. Tupin, F., Maitre, H., Mangin, J.-F., Nicolas, J.-M., Pechersky, E.: Detection of linear features in SAR images: application to road network extraction. *IEEE Trans. Geosci. Remote Sens.* **36**, 434–453 (1998)
37. Zomorodian, A., Carlsson, G.: Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005)