Metric Graph Reconstruction from Noisy Data

Mridul Aanjaneya
Stanford University
Stanford, California 94305
aanjneya@stanford.edu

Frederic Chazal
INRIA Saclay – Île-de-France
Orsay, France
frederic.chazal@inria.fr

Daniel Chen
Stanford University
Stanford, California 94305
danielc@cs.stanford.edu

Marc Glisse
INRIA Saclay – Île-de-France
Orsay, France
marc.glisse@inria.fr

Leonidas Guibas
Stanford University
Stanford, California 94305
guibas@cs.stanford.edu

Dmitriy Morozov
Stanford University
Stanford, California 94305
dmitriy@mrzv.org

ABSTRACT

Many real-world data sets can be viewed of as noisy samples of special types of metric spaces called *metric graphs* [16]. Building on the notions of correspondence and Gromov-Hausdorff distance in metric geometry, we describe a model for such data sets as an approximation of an underlying metric graph. We present a novel algorithm that takes as an input such a data set, and outputs the underlying metric graph with guarantees. We also implement the algorithm, and evaluate its performance on a variety of real world data sets.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*geometric*; F.2.2 [**Nonnumerical Algorithms and Problems**]: [geometrical problems and computations]

General Terms

Algorithms, Experimentation

Keywords

Reconstruction, metric graph, noise, inference

1. INTRODUCTION

Motivation

Large-scale geometric data sets are becoming widely available, whether from high-bandwidth sensors or from massive simulations of physical processes. All across science, engineering, medicine, and defense, there is a real need to analyze, understand, and extract useful information out of such massive geometric data. Much of this data is noisy, contains outliers, has missing parts, and does not have a manifold structure or even a consistent dimension — raising many difficult statistical, geometric, and algorithmic problems in its analysis. In this paper, we focus on a simple, but important setting of mixed-dimension geometric data, namely a setting where the underlying space of the data can be viewed of as a *metric graph* [16], which is an 1-D stratified space consisting of just 0-D strata (vertices) and 1-D linear strata (edges or loops), glued together in some fashion, see Figure 1(a).

Branching filamentary structures, which can be naturally viewed of as metric graphs, appear in a wide variety of real-world data sets, both in settings where the data arises embedded in Euclidean space, as well as in situations where the host space is less intuitive and only local metric information may be available. For example large-scale collections of GPS traces for vehicles or pedestrians are becoming widely available (see e.g., [2]) and can be used to provide a variety of location-aware services. Their movement patterns tend to follow a branching structure which can be modeled as a metric graph. Earthquake faults are intimately connected with plate tectonics and tend to follow filamentary structures as they arise along the boundaries of such plates (see e.g., [1]). In nuclear physics, high-energy particles move along filamentary trajectories and there is often the need to track their motion [3]. In materials science, stresses can cause material cracks that propagate along branching structures formed by linear paths; their detection is an important research problem [17]. Many defense applications require the extraction of road networks from synthetic aperture radar (SAR) images [19]. In astronomy, filamentary structures in galaxies are of great interest (e.g., [9]) for cosmological studies. This is not to mention networks formed by blood vessels in the body for anatomy, river systems in geography, and many other examples.

Branching structures are also quite common in more abstract settings, though sometimes one has to look at such data with a coarser lens before it becomes apparent. For instance, communication networks can be regarded as large graphs in which certain dominant pathways define the major arteries connecting network hubs. Recently, Heath *et. al.* [15] built large graphs from image collections, by linking together images with partial shared content. Data sets of interest here include collections of images acquired by a mobile agent along its path, as in Google Streetview. In such cases, at a coarse scale, the connectivity among the images reflects the mobility of the capturing agent, naturally giving rise to branching filamentary structures. Extraction of this under-

SCG'11, June 13–15, 2011, Paris, France.

lying structure can provide a useful map for understanding the image data, navigating through it, or for answering certain queries.

Reconstruction Problem

While there has been a great deal of prior work on both topological and geometric reconstruction of geometric data sets under varying sampling conditions, our emphasis is on an intermediate level of reconstruction, what we term *metric reconstruction* — a largely unexplored domain. The input to our algorithm is a metric space $(\mathbb{Y}, d_{\mathbb{Y}})$ that is close to a much simpler metric graph $(\mathbb{X}, d_{\mathbb{X}})$ in a sense that we make precise in the Section 3. $(\mathbb{Y}, d_{\mathbb{Y}})$ can be constructed from raw data in various ways: in some cases, we construct a neighborhood graph on the raw data, and use the shortest path as the distance; in other cases, the metric is given to us directly. Note that this implies that our reconstruction is aimed at capturing the intrinsic structure of the data and is somewhat oblivious to its extrinsic embedding, wherever that is available. Our goal is then to extract a metric graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ that has the same topology as $(\mathbb{X}, d_{\mathbb{X}})$, and a map $\phi : \mathbb{Y} \to \hat{\mathbb{X}}$ that approximately preserves distances.

Experiments

In addition to theoretical reconstruction results with performance guarantees, we study experimentally the performance of our algorithm on a variety of data sets from different applications, including data in which an embedding is given (GPS traces, earthquake data, astronomical data), as well as data in which only metric information is available (Image Webs). In all these cases our compact metric approximation provides a much more manageable representation of the structure of the original data — far easier to visualize, navigate, and manipulate than the original. Our metric guarantees allow us to further exploit this representation by running graph algorithms in this compact representation in lieu of the original graph. As an example, we used the compressed graph to perform shortest path queries, resulting in significant speedups on some data sets.

Related Work

Our work is related to contributions by several different communities. On the one side, the statistics community has investigated the problem of extracting filamentary structures from point cloud data, starting with the seminal work of Arias-Castro *et al.* [5] based on counting membership in multiscale anisotropic strips. Subsequent approaches exploit gradient descent or medial axis ideas [12, 13]. All these, however, aim mostly at the extraction of isolated filaments, focus on how to deal with outlier data, and do not pay serious attention to the global branching structures the filaments form. Also, they all assume an extrinsic embedding of the data. On the other side, there has been extensive work in the computational geometry community on curve reconstruction, which is the problem of computing a polygonal curve that approximates well a curve sampled by a given point set — several algorithms have been proposed for this problem [4, 10, 11]. Unfortunately, it is hard to extend these methods to our setting, since they also view 0-dimensional strata, which exhibit non-manifold behavior, as singularities and try to avoid them as much as possible. While geometric reconstruction is not our goal, as in that work, we aim to be able to prove certain quality guarantees on the metric reconstruction we attain, under appropriate sampling conditions. Finally, Chen *et al.* [8] recently considered a related problem of reconstructing a road network from a given collection of path traces. They designed an algorithm with guarantees without making heavy assumptions on the distribution of input paths. However, the assumptions they use are stronger than desired in many practical applications. In particular, their method depends on an embedding of the data and sequential path information.

We end by remarking that dimension reduction has been a topic of much study in the machine learning and data analysis communities. When data is given in parametric form, i.e., as points in a (possibly high dimensional) Euclidean space and the goal of dimension reduction is distance preservation, many well-known methods exist based on random projections as suggested by the Johnson-Lindenstrauss lemma, or by locality sensitive hashing (LSH). This paper addresses "dimension reduction" for distance preservation in the case where the metric is given by the shortest path distance on a large but special type of graph — one that contains few but large linear structures. As we show, this type of metric reconstruction raises interesting new mathematical problems and is applicable to many types of geometric data.

2. PRELIMINARIES

Recall that a *metric space* is a pair $(\mathbb{X}, d)$ where $\mathbb{X}$ is a set and $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ is a symmetric function satisfying (1): $d(x, x') = 0$ if and only if $x = x'$ and (2): $d(x, x'') \leq d(x, x') + d(x', x'')$. Two spaces $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ are *isometric* if there exists a bijection $\phi : \mathbb{X} \to \mathbb{Y}$ that preserves the distances, namely: $d_{\mathbb{Y}}(\phi(x), \phi(x')) = d_{\mathbb{X}}(x, x')$ for all $x, x' \in \mathbb{X}$. The space of isometry classes of metric spaces is endowed with the *Gromov-Hausdorff distance* [14] whose definition can be given using the notion of ε-correspondences ([6] Thms 7.3.25 and 7.3.30).

Definition 1. A *correspondence* between $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ is a set $C \subset \mathbb{X} \times \mathbb{Y}$ such that for any $x \in \mathbb{X}$ (resp. $y \in \mathbb{Y}$), there exists $y \in \mathbb{Y}$ (resp. $x \in \mathbb{X}$) such that $(x, y) \in C$. When x, y are such that $(x, y) \in C$, we say that x and y are paired in C. Given $\varepsilon > 0$, C is an *ε-correspondence* if for any $(x, y), (x', y') \in C$, $|d_{\mathbb{X}}(x, x') - d_{\mathbb{Y}}(y, y')| \leq \varepsilon$. The *Gromov-Hausdorff distance* $d_{GH}(\mathbb{X}, \mathbb{Y})$ is the infimum of the $\varepsilon \geq 0$ such that there exists an ε-correspondence between $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$.

An ε-correspondence between $\mathbb{X}$ and $\mathbb{Y}$ can be seen as an ε-approximation of $\mathbb{X}$ by $\mathbb{Y}$ (and reciprocally). However, in many applications, data only comes with locally correct approximate metric information. For example, for a data set sampling a road network the Euclidean distance between data points provides a suitable approximation of the metric of the underlying network only locally. So in this paper, we use a more local and weaker notion of correspondence: given positive numbers ε, R, we say that $(\mathbb{Y}, d_{\mathbb{Y}})$ is an (ε, R)-*approximation* of a metric space $(\mathbb{X}, d_{\mathbb{X}})$ if there exists a correspondence $C \subset \mathbb{X} \times \mathbb{Y}$ such that

$$(x, y), (x', y') \in C,\ \min(d_{\mathbb{X}}(x, x'), d_{\mathbb{Y}}(y, y')) \leq R$$

$$\implies |d_{\mathbb{X}}(x, x') - d_{\mathbb{Y}}(y, y')| \leq \varepsilon$$

Notice that this latter notion is strictly weaker than the notion of ε-correspondence. In particular, the existence of

an (ε, R)-correspondence between $\mathbb{X}$ and $\mathbb{Y}$ does not bound $d_{GH}(\mathbb{X}, \mathbb{Y})$, as shown in the following example: let $\mathbb{X} \subset \mathbb{R}^2$ be the half-circle $\{x^2 + y^2 = 1, y \leq 0\}$ endowed with the geodesic distance and let $\mathbb{Y} = \mathbb{X}$ be endowed with the restriction of the Euclidean distance. For any $\varepsilon > 0$, the diagonal $C = \{(x, x) : x \in \mathbb{X}\} \subset \mathbb{X} \times \mathbb{Y}$ is an $(\varepsilon, O(\varepsilon^{1/3}))$-correspondence, but the diameters of $\mathbb{X}$ and $\mathbb{Y}$ are respectively equal to π and 2, showing that $d_{GH}(\mathbb{X}, \mathbb{Y}) \geq \pi - 2 > 0$. Nevertheless, (ε, R)-approximations give rise to global approximations with respect to d_{GH} when the approximated space is a *path metric space*, defined as follows:

Definition 2. A metric space $(\mathbb{X}, d_{\mathbb{X}})$ is a *path metric space* if the distance between any pair of points is equal to the infimum of the lengths of the continuous curves joining them [1]. Equivalently $(\mathbb{X}, d_{\mathbb{X}})$ is a path metric space if and only if for any $x, y \in \mathbb{X}$ and any $\varepsilon > 0$ there exists $z \in \mathbb{X}$ such that $\max(d_{\mathbb{X}}(x, z), d_{\mathbb{X}}(y, z)) \leq \frac{1}{2} d_{\mathbb{X}}(x, y) + \varepsilon$ [14].

Then, we can obtain the following bound on the Gromov-Hausdorff distance:

LEMMA 1. *Let $(\mathbb{X}, d_{\mathbb{X}})$ be a path metric space, $(\mathbb{Y}, d_{\mathbb{Y}})$ an (ε, R)-approximation of $\mathbb{X}$ and assume that $\mathbb{Y}$ has the following property:*

($\star$) for any $y, y' \in \mathbb{Y}$ there exists a sequence $y_0 = y, y_1, \ldots y_{n-1}, y_n = y'$ such that for all $i = 0, \ldots, n-1$, $d_{\mathbb{Y}}(y_i, y_{i+1}) \leq R$ and $d_{\mathbb{Y}}(y, y') = \sum_{i=0}^{n-1} d_{\mathbb{Y}}(y_i, y_{i+1})$.

If $C \subset \mathbb{X} \times \mathbb{Y}$ is an (ε, R)-correspondence, then for any $(x, y), (x', y') \in C$ we have

$$|d_{\mathbb{Y}}(y, y') - d_{\mathbb{X}}(x, x')| \leq \left(\frac{\min(d_{\mathbb{X}}(x, x'), d_{\mathbb{Y}}(y, y'))}{R/2} + 1 \right) \varepsilon.$$

In particular, $d_{GH}((\mathbb{X}, d_{\mathbb{X}}), (\mathbb{Y}, d_{\mathbb{Y}})) \leq \left(\frac{\mathrm{diam}(\mathbb{X})}{R/2} + 1 \right) \varepsilon$ *where* $\mathrm{diam}(\mathbb{X})$ *is the diameter of* $\mathbb{X}$.

PROOF. (x, y) and $(x', y') \in C$ are given. By hypothesis, there exists a sequence $y_0 = y, y_1, \ldots y_{n-1}, y_n = y'$ such that for all $i = 0, \ldots, n-1$, $d_{\mathbb{Y}}(y_i, y_{i+1}) \leq R$ and $d_{\mathbb{Y}}(y, y') = \sum_{i=0}^{n-1} d_{\mathbb{Y}}(y_i, y_{i+1})$.

As a first remark, notice that in this sequence, if $i < j$, $d_{\mathbb{Y}}(y_i, y_j) = \sum_{k=i}^{j-1} d_{\mathbb{Y}}(y_k, y_{k+1})$. Indeed, using the triangle inequality: $d_{\mathbb{Y}}(y, y') \leq \sum_{k=0}^{i-1} d_{\mathbb{Y}}(y_k, y_{k+1}) + d_{\mathbb{Y}}(y_i, y_j) + \sum_{k=j}^{n-1} d_{\mathbb{Y}}(y_k, y_{k+1}) \leq \sum_{k=0}^{n-1} d_{\mathbb{Y}}(y_k, y_{k+1}) = d_{\mathbb{Y}}(y, y')$.

In property ($\star$), we can further assume that $d_{\mathbb{Y}}(y_i, y_{i+2}) > R$. If $d_{\mathbb{Y}}(y_i, y_{i+2}) \leq R$, we can remove y_{i+1} from the sequence, and the previous remark shows that the properties are still satisfied. In particular, this implies that $d_{\mathbb{Y}}(y, y') > \frac{n-1}{2} R$.

Now to each y_i corresponds a (non-unique) $x_i \in \mathbb{X}$ in C.

$$d_{\mathbb{X}}(x, x') \leq \sum_{i=0}^{n-1} d_{\mathbb{X}}(x_k, x_{k+1}) \leq \sum_{i=0}^{n-1} d_{\mathbb{Y}}(y_k, y_{k+1}) + n\varepsilon \leq d_{\mathbb{Y}}(y, y') + \left(\frac{d_{\mathbb{Y}}(y, y')}{R/2} + 1 \right) \varepsilon.$$

A simple computation shows that this implies:

$$d_{\mathbb{X}}(x, x') < d_{\mathbb{Y}}(y, y') + \left(\frac{d_{\mathbb{X}}(x, x')}{R/2} + 1 \right) \varepsilon.$$

[1] see [14] Chap.1 for the definition of the length of a continuous curve in a general metric space

Now $\mathbb{X}$ almost satisfies ($\star$). Indeed, by recursively splitting the intervals of length more that R, for any $\varepsilon' > 0$, we construct a sequence $x_0 = x, x_1, \ldots, x_n = x'$ such that $d_{\mathbb{X}}(x_i, x_{i+1}) \leq R$, $d_{\mathbb{X}}(x_i, x_{i+2}) \geq R - \varepsilon'$ and $\sum_{i=0}^{n-1} d_{\mathbb{X}}(x_i, x_{i+1}) \leq d_{\mathbb{Y}}(y, y') + \varepsilon'$. We derive as before:

$$d_{\mathbb{Y}}(y, y') \leq d_{\mathbb{X}}(x, x') + \left(\frac{d_{\mathbb{X}}(x, x') + \varepsilon'}{R/2 - \varepsilon'} + 1 \right) \varepsilon$$

and since it is true for all $\varepsilon' > 0$:

$$d_{\mathbb{Y}}(y, y') \leq d_{\mathbb{X}}(x, x') + \left(\frac{d_{\mathbb{X}}(x, x')}{R/2} + 1 \right) \varepsilon$$

which again implies:

$$d_{\mathbb{Y}}(y, y') < d_{\mathbb{X}}(x, x') + \left(\frac{d_{\mathbb{Y}}(y, y')}{R/2} + 1 \right) \varepsilon$$

and finally:

$$|d_{\mathbb{Y}}(y, y') - d_{\mathbb{X}}(x, x')| \leq \left(\frac{\min(d_{\mathbb{X}}(x, x'), d_{\mathbb{Y}}(y, y'))}{R/2} + 1 \right) \varepsilon$$

□

In this paper, we assume that our input is an (ε, R)-approximation of a specific type of path metric space, known as a *metric graph* [16]:

Definition 3. A *metric graph* is a path metric space $(\mathbb{X}, d_{\mathbb{X}})$ that is homeomorphic to a 1-dimensional stratified space (see Figure 1(a)). A *vertex* of $\mathbb{X}$ is a 0-dimensional stratum of $\mathbb{X}$ and an *edge* of $\mathbb{X}$ is a 1-dimensional stratum of $\mathbb{X}$ [2].

It is useful to note that edges are isometric to finite length intervals in the real line.

3. PROBLEM DEFINITION

Let $(\mathbb{Y}, d_{\mathbb{Y}})$ be an (ε, R)-approximation of a metric graph $(\mathbb{X}, d_{\mathbb{X}})$ that has a shortest edge length of b. Without loss of generality, we will assume that $\mathbb{X}$ is connected. Note that our definition of (ε, R)-approximation is essentially a worst-case noise model for the data that does not rely on further distributional assumptions. In practice, such a $(\mathbb{Y}, d_{\mathbb{Y}})$ is often obtained by building a (weighted) neighborhood graph on a raw data set $\mathbb{Y}$, and defining $d_{\mathbb{Y}}(y_1, y_2)$ to be the length of the shortest path joining y_1 and y_2 on the graph $\forall y_1, y_2 \in \mathbb{Y}$. Additionally, we assume that $(\mathbb{Y}, d_{\mathbb{Y}})$ satisfies the property ($\star$) of Lemma 1, and if this property is not satisfied, we can instead consider the so-called *Rips-Vietoris graph* $\mathcal{R}_R(\mathbb{Y})$ with vertex set $\mathbb{Y}$ and edges connecting all the pairs of vertices at distance less than R from each other in $\mathbb{Y}$. The metric $\tilde{d}_{\mathbb{Y}}$ induced by this graph coincides with $d_{\mathbb{Y}}$ for the pairs of points at distance less than R and therefore $(\mathbb{Y}, \tilde{d}_{\mathbb{Y}})$ is still an (ε, R)-approximation of $(\mathbb{X}, d_{\mathbb{X}})$. Our goal is to design an algorithm to reconstruct from $(\mathbb{Y}, d_{\mathbb{Y}})$ a space $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ that is homeomorphic to $(\mathbb{X}, d_{\mathbb{X}})$. Furthermore, we define distances on $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ that approximate those of $(\mathbb{X}, d_{\mathbb{X}})$ and return a map $\phi : \mathbb{Y} \to \hat{\mathbb{X}}$ that approximately preserves distances. Although we frame this objective as a reconstruction problem, in practice, our algorithm can be used to find a much simpler metric graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ approximating the input space $(\mathbb{Y}, d_{\mathbb{Y}})$, and achieving guarantees when $(\mathbb{Y}, d_{\mathbb{Y}})$ is an approximation of a suitable metric graph $(\mathbb{X}, d_{\mathbb{X}})$.

[2] We also include in our definition the 1-dimensional manifold isometric to a circle (one edge and no vertex)

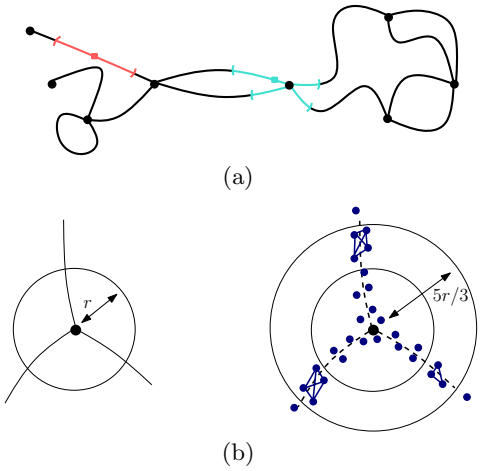

Figure 1: (a) A metric graph (in black) and 2 intrinsic balls (in blue and red). (b) Using a spherical shell to infer the degree of a vertex.

4. OVERVIEW OF ALGORITHM AND GUARANTEES

In addition to the input metric $(\mathbb{Y}, d_{\mathbb{Y}})$, which, as mentioned previously, is an (ε, R)-approximation of an underlying metric graph $(\mathbb{X}, d_{\mathbb{X}})$, our algorithm also takes a parameter r that roughly corresponds to the scale at which we look at the data. For noisier data, we would generally use a larger r, while to capture smaller features, we would choose a smaller r. Our analysis will exhibit a range of values for r that result in a correct reconstruction depending on both ε and R, as well as on b, the length of the shortest edge in $\mathbb{X}$. In practice, we do not know these values, but our implementation always outputs a suitable metric graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ for which we can check the distortion of the metric from $(\mathbb{Y}, d_{\mathbb{Y}})$. Hence, we are able to try values of r until we obtain a suitable and simple approximation of $(\mathbb{Y}, d_{\mathbb{Y}})$.

Recall that there is an (ε, R)-correspondence between our input metric $(\mathbb{Y}, d_{\mathbb{Y}})$ and its underlying metric graph $(\mathbb{X}, d_{\mathbb{X}})$. The algorithm proceeds in two steps. First it begins by labeling as "branch points" the points of $\mathbb{Y}$ paired under this correspondence to a point in $\mathbb{X}$ that is close to a vertex and labeling the rest of the points of $\mathbb{Y}$ as "edge points". Then, the algorithm uses these labels to reconstruct a new metric graph $\hat{\mathbb{X}}$ that is homeomorphic to $\mathbb{X}$ and estimates distance preserving maps from $\mathbb{Y}$ to $\hat{\mathbb{X}}$. For ease of reference, the pseudocode of our algorithm is given in Algorithm 1.

The following results show that if $(\mathbb{Y}, d_{\mathbb{Y}})$ is a sufficiently good approximation of $(\mathbb{X}, d_{\mathbb{X}})$ then the reconstructed graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ is homeomorphic and almost isometric to $(\mathbb{X}, d_{\mathbb{X}})$.

THEOREM 1 (TOPOLOGICAL RECONSTRUCTION). *If the length b of the shortest edge of $\mathbb{X}$ is larger than $16r$ and $15\varepsilon/2 < r < \min(R/4, 3(b-2\varepsilon)/5)$ then the reconstructed graph $\hat{\mathbb{X}}$ is homeomorphic to $\mathbb{X}$.*

THEOREM 2 (METRIC RECONSTRUCTION). *Under the assumptions of Theorem 1 there exists a homeomorphism $\phi : \mathbb{X} \to \hat{\mathbb{X}}$ such that for any $x, x' \in \mathbb{X}$, $(1-\kappa) d_{\mathbb{X}}(x, x') \leq d_{\hat{\mathbb{X}}}(\phi(x), \phi(x')) \leq (1+\kappa') d_{\mathbb{X}}(x, x')$ with $\kappa = \frac{10r}{3b} + (\frac{5}{b} + \frac{2}{R})\varepsilon$ and $\kappa' = (\frac{3}{b} + \frac{2}{R})\varepsilon$.*

Algorithm 1 Metric Graph Reconstruction

Require: Metric space $(\mathbb{Y}, d_{\mathbb{Y}})$ approximating metric graph $(\mathbb{X}, d_{\mathbb{X}})$ and parameter $r > 0$.
Ensure: Metric graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$

```
LABELING POINTS AS EDGE OR BRANCH
for all y ∈ Y do
    S_y ← B_Y(y, 5r/3) \ B_Y(y, r))
    deg_r(y) ← Number of connected components of Rips-Vietoris graph R_{4r/3}(S_y)
    if deg_r(y) = 2 then
        Label y as a edge point.
    else
        Label y as a preliminary branch point.
    end if
end for
Label all points within distance 2r from a preliminary branch point as branch points.
Let E be the points of Y labeled as edge points.
Let V be the points of Y labeled as branch points.
RECONSTRUCTING THE GRAPH STRUCTURE
Compute the connected components of the Rips-Vietoris graphs R_{2r}(E) and R_{2r}(V).
Let the connected components of R_{2r}(V) be the vertices of the reconstructed graph X̂.
Let there be an edge between vertices of X̂ if their corresponding connected components in R_{2r}(V) contain points at distance less than 2r from the same connected component of R_{2r}(E).
RECONSTRUCTING THE METRIC
To each edge ê of X̂ assign a length equal to the diameter of the corresponding connected component of R_{2r}(E) plus 4r.
```

The proofs of these results, as well as a more detailed discussion of the algorithm, are given in the next section where an easy to compute map with low-metric distortion between $(\mathbb{Y}, d_{\mathbb{Y}})$ and $(\mathbb{X}, d_{\hat{\mathbb{X}}})$ is also provided.

5. ANALYSIS AND PROOFS

In this section we assume that the assumptions of Theorems 1 and 2 are satisfied.

5.1 Labeling points as edge or branch

First notice that the classification of a point $x \in \mathbb{X}$ as a vertex or a point on an edge is determined by the number of connected components of a small intrinsic sphere centered at x (see Figure 1(b)). To label a point $y \in \mathbb{Y}$ as either a branch point or an edge point, our algorithm considers the intrinsic spherical shells $\mathrm{B}_{\mathbb{Y}}(y, 5r/3) \setminus \mathrm{B}_{\mathbb{Y}}(y, r)$ around y and constructs a Rips-Vietoris graph with parameter $4r/3$ on the points of $\mathbb{Y}$ inside the spherical shell. Then, it records the number of connected components of this graph as the *r-degree* $deg_r(y)$:

Definition 4. Let $(\mathbb{Y}, d_{\mathbb{Y}})$ be an (ε, R)-approximation of $\mathbb{X}$. Given $0 < r < R/2$, the *r-degree* $deg_r(y)$ of a point $y \in \mathbb{Y}$ is the number of connected components of the Rips-Vietoris graph $\mathcal{R}_{4r/3}(\mathrm{B}_{\mathbb{Y}}(y, 5r/3) \setminus \mathrm{B}_{\mathbb{Y}}(y, r))$ with vertex set $\mathrm{B}_{\mathbb{Y}}(y, 5r/3) \setminus \mathrm{B}_{\mathbb{Y}}(y, r)$ and edges connecting all the pairs of vertices at distance less than $4r/3$ from each other.

Intuitively, it is easy to imagine that if $deg_r(y) \neq 2$, then y corresponds to a point on $\mathbb{X}$ close to a vertex, whereas if $deg_r(y) = 2$, y corresponds to a point on $\mathbb{X}$ far from a vertex.

THEOREM 3 (DEGREE INFERENCE THEOREM). *Let $(\mathbb{Y}, d_{\mathbb{Y}})$ be an (ε, R)-approximation of $\mathbb{X}$. Let $C \subset \mathbb{X} \times \mathbb{Y}$*

be an (ε, R)-correspondence between $\mathbb{X}$ and $\mathbb{Y}$, let $(x, y) \in C$.
i) If the distance d_0 from x to any vertex of $\mathbb{X}$ is larger than $\frac{17}{2}\varepsilon$, then for $\frac{9}{2}\varepsilon < r < \min(\frac{R}{2}, \frac{3(d_0-\varepsilon)}{5})$, $deg_r(y)$ is equal to the degree of x in $\mathbb{X}$ (i.e. 2). Moreover the pairwise distances between the connected components of the Rips-Vietoris graph are lower bounded by $2r - 3\varepsilon$.
ii) If x is at distance less than ε from a vertex x_0 of $\mathbb{X}$ and if the length l_0 of the shortest edge adjacent to x_0 is larger than $\frac{27}{2}\varepsilon$ then for $\frac{15}{2}\varepsilon < r < \min(\frac{R}{2}, \frac{3(l_0-2\varepsilon)}{5})$, $deg_r(y)$ is equal to the degree of x_0 in $\mathbb{X}$. Moreover the pairwise distances between the connected components of the Rips-Vietoris graph are lower bounded by $2r - 5\varepsilon$.

PROOF. This theorem is a consequence of Theorem 5 in Appendix A with α set to $2/3$. □

In Appendix A, we consider a more general variation of the r-degree dependent on an extra parameter $0 \leq \alpha \leq 1$, which allows us to vary the radius of the ball $\mathrm{B}_{\mathbb{Y}}(y, 5r/3)$ to any $(1+\alpha)r$ for different guarantees. Nevertheless, choosing $\alpha = 2/3$ optimizes the number of points that are "correctly" inferred as edge points (see Appendix A).

5.2 Reconstructing the graph structure

We now describe the reconstruction procedure. Given $15\varepsilon/2 < r < \min(R/4, 3(b-2\varepsilon)/5)$, recall that we first label the points $y \in \mathbb{Y}$ as *branch* or *edge* depending on $deg_r(y)$: y is labelled as an edge point if $deg_r(y) = 2$, and labelled as a branch point otherwise. The following result is an immediate consequence of Theorem 3.

LEMMA 2. *If $y \in \mathbb{Y}$ is paired in C to a point x at distance at most ε from a vertex of $\mathbb{X}$ then y is labeled as a branch point by the procedure above. If $y \in \mathbb{Y}$ is paired in C to a point x at distance at least $5r/3 + \varepsilon$ from any vertex of $\mathbb{X}$ then y is labeled as an edge point.*

The points of $\mathbb{Y}$ paired to points in $\mathbb{X}$ that are at distance between ε and $5r/3+\varepsilon$ from a vertex of $\mathbb{X}$ can be "incorrectly" labeled as branch points. It is not possible to distinguish these *fuzzy* points from the data $\mathbb{Y}$ only, so we force them to be branch points using the following expansion procedure: all points $y \in \mathbb{Y}$ that are at distance at most $2r$ from a point labeled as branch are promoted to branch points.

To prove that after this expansion all the fuzzy points are labeled as branch, notice that if $y \in \mathbb{Y}$ is now labeled as an edge then it is at distance at least $2r$ from any point $y' \in \mathbb{Y}$ labeled as branch before the expansion procedure. It follows that for any pair $(x, y) \in C$, x is at distance more than $2r - \varepsilon > 5r/3 + \varepsilon$ (since $r > 15\varepsilon/2$) from a vertex of $\mathbb{X}$.

COROLLARY 1. *Let (x, y) be a pair in C. If x is at distance at least $11r/3+2\varepsilon$ from any vertex of $\mathbb{X}$, then after the expansion procedure, y is labeled as an edge. Reciprocally, if y is labeled as an edge after the expansion procedure, then x is at distance at least $2r - \varepsilon$ from a vertex of $\mathbb{X}$.*

Now to recover the connectivity of $\mathbb{X}$, we group the branch points (resp. the edge points) in clusters, each corresponding to a vertex (resp. an edge) of $\mathbb{X}$. For that, we consider the Rips-Vietoris graph $\mathcal{R}_{2r}(\mathbb{V})$ (resp. $\mathcal{R}_{2r}(\mathbb{E})$) of parameter $2r$ built on top of the set $\mathbb{V} \subset \mathbb{Y}$ of branch points (resp. the set $\mathbb{E} \subset \mathbb{Y}$ of edge points).

LEMMA 3. *If the length b of the shortest edge of $\mathbb{X}$ is larger than $16r$ then the connected components of $\mathcal{R}_{2r}(\mathbb{V})$ are in one-to-one correspondence with the vertices of $\mathbb{X}$ and the connected components of $\mathcal{R}_{2r}(\mathbb{E})$ are in one-to-one correspondence with the edges of $\mathbb{X}$.*

PROOF. If $y \in \mathbb{Y}$ is a branch point and $(x, y) \in C$, then there exist $(x', y'), (x_0, y_0) \in \mathbb{X}$ such that x_0 is a vertex in $\mathbb{X}$, $d_{\mathbb{X}}(x_0, x') \leq 5r/3 + \varepsilon$ and $d_{\mathbb{X}}(y', y) \leq 2r$. It follows that $d_{\mathbb{Y}}(y_0, y') \leq 5r/3 + 2\varepsilon \leq 2r$. So y and y_0 are in the same connected component of $\mathcal{R}_{2r}(\mathbb{V})$ and $d_{\mathbb{Y}}(y, y_0) \leq 4r$. As a consequence, any connected component of $\mathcal{R}_{2r}(\mathbb{V})$ contains at least one point paired with a vertex of $\mathbb{X}$.

Now if $(x_1, y_1) \in C$ is such that x_1 is another vertex of $\mathbb{X}$, then Lemma 1 implies that

$$d_{\mathbb{Y}}(y_0, y_1) \geq d_{\mathbb{X}}(x_0, x_1) - (\frac{2d_{\mathbb{X}}(x_0, x_1)}{R} + 1)\varepsilon \geq \frac{4}{5}b$$

where to get the last inequality we used that $d_{\mathbb{X}}(x_0, x_1) \geq b$, $R > 15\varepsilon$ and $b > 15\varepsilon$. Assume that y_0 and y_1 are in the same connected component of $\mathcal{R}_{2r}(\mathbb{V})$. Then there exists a path joining y_0 to y_1 in this component and since $x_0 \neq x_1$, there exists a branch point $y' \in \mathbb{Y}$ along this path such that $b/2 - r \leq d_{\mathbb{Y}}(y', y_0) \leq b/2 + r$. According to Lemma 1 for any $x' \in \mathbb{X}$ such that $(x', y') \in C$, we have

$$d_{\mathbb{Y}}(y', y_0)(1 - \frac{2\varepsilon}{R}) - \varepsilon \leq d_{\mathbb{X}}(x', x_0) \leq d_{\mathbb{Y}}(y', y_0)(1 + \frac{2\varepsilon}{R}) + \varepsilon$$

Using again that $\varepsilon/R < 1/15$ and $\varepsilon < b/15$ we get $\frac{11}{30}b - \frac{13}{15}r \leq d_{\mathbb{X}}(x', x_0) \leq \frac{19}{30}b + \frac{17}{15}r$ and since b is the length of the shortest edge of $\mathbb{X}$, the distance between x' and any vertex of $\mathbb{X}$ is at least $\min(\frac{11}{30}b - \frac{13}{15}r, b - (\frac{19}{30}b + \frac{17}{15}r)) = \frac{11}{30}b - \frac{17}{15}r$. Since $b > 16r$, one deduces from the corollary 1 that x' is an edge point: a contradiction. As a consequence, the points of any connected component of $\mathcal{R}_{2r}(\mathbb{V})$ can be paired with at most one vertex of $\mathbb{X}$. This proves that the connected components of $\mathcal{R}_{2r}(\mathbb{V})$ are in one-to-one correspondence with the vertices of $\mathbb{X}$.

To prove the second part of the lemma, first notice that since $b > 16r$ for any edge of $\mathbb{X}$ there exists a point at distance at least $8r$ from any vertex of $\mathbb{X}$. As a consequence, any $y \in \mathbb{Y}$ such that $(x, y) \in C$ is labeled as an edge point showing that $\mathbb{E}$ contains points from all the edges of $\mathbb{X}$. Now if $(x, y), (x', y') \in C$ are such that $y, y' \in \mathbb{E}$ and x, x' are not in the same edge of $\mathbb{X}$, then any shortest path joining x to x' has to meet a vertex x'' of $\mathbb{X}$. So for any sequence $(x_0, y_0) = (x, y), (x_1, y_1), \cdots (x_n, y_n) = (x', y') \in C$ such that $y_0 = y, y_1 \ldots, y_n = y'$ is joining y to y' in $\mathcal{R}_{2r}(\mathbb{Y})$ there exists $i \in \{1, \cdots, n-1\}$ such that $d_{\mathbb{X}}(x'', x_i) \leq \frac{1}{2}(2r + \varepsilon)$. It follows that y_i is a branch point and y and y' cannot be in the same connected component of $\mathcal{R}_{2r}(\mathbb{E})$. Reciprocally, if x, x' are in the same edge e of $\mathbb{X}$, they both are at distance at least $2r - \varepsilon$ from the end points of e and from any point paired to a point labeled as branch before the expansion procedure. So, if $(x'', y'') \in C$ is such that $x'' \in e$ is contained in the interval defined by x and x' and is at distance larger than $2r - \varepsilon$ from x and x', then the distance from x'' to any point paired to a branch point before the expansion procedure is at least $4r - 2\varepsilon > 11r/3 + 2\varepsilon$ (since $r > 15\varepsilon/2$). Therefore, y'' is an edge point. As a consequence, there exists a sequence $y_0 = y, y_1 \ldots, y_n = y'$ of edge points that are all paired to points in the edge e

such that $d_{\mathbb{Y}}(y_i, y_{i+1}) \leq 2r$ for $i = 0, \dots n-1$, proving that y and y' are in the same connected component of $\mathcal{R}_{2r}(\mathbb{E})$. It follows that the connected components of $\mathcal{R}_{2r}(\mathbb{E})$ are in one-to-one correspondence with the edges of $\mathbb{X}$.

□

Now recall that $\hat{\mathbb{X}}$ is built as follows: we create a vertex for each connected component of $\mathcal{R}_{2r}(\mathbb{V})$; we create an edge between two vertices if each of the two corresponding components contains at least one point at distance less than $2r$ from the same connected components of $\mathcal{R}_{2r}(\mathbb{E})$. From Lemma 3 we then deduce the Topological Reconstruction Theorem 1.

5.3 Reconstructing the metric

We begin with the proof of Theorem 2:

PROOF OF THEOREM 2. The proof consists of showing the existence of a $(1+\kappa')$-Lipschitz homeomorphism $\phi : \mathbb{X} \to \hat{\mathbb{X}}$ with inverse $(1-\kappa)^{-1}$-Lipschitz. To this end, we proceed with each edge separately. Let $\hat{e}$ be an edge of $\hat{\mathbb{X}}$, and let y_0, y_1 be two points in the corresponding connected component in $\mathcal{R}_{2r}(\mathbb{E})$ such that $d_{\mathbb{Y}}(y_0, y_1)$ is equal to the diameter of this component. Denoting e the edge of $\mathbb{X}$ corresponding to $\hat{e}$, Corollary 1 implies that y_0 and y_1 are paired in C to points in e that are located at distance at least $2r - \varepsilon$ from the extremities of e. As a consequence of Lemma 1 we have $d_{\mathbb{Y}}(y_0, y_1) \leq (1 + 2\varepsilon/R) l(e) - 4r + 3\varepsilon$. Now, let $(x, y), (x', y') \in C$ such that $x, x' \in e$ are two points at distance $11r/3 + 2\varepsilon$ from each endpoint of e. We deduce from Corollary 1 that y, y' are edge points, so $d_{\mathbb{Y}}(y, y') \leq d_{\mathbb{Y}}(y_0, y_1)$ and from Lemma 1 that

$$\begin{aligned} d_{\mathbb{Y}}(y, y') &\geq d_{\mathbb{X}}(x, x') - (2 d_{\mathbb{X}}(x, x')/R + 1)\varepsilon \\ &\geq l(e)\left(1 - \frac{2\varepsilon}{R}\right) - \frac{22r}{3} - 5\varepsilon \end{aligned}$$

where for the last inequality we have used that $l(e) = d_{\mathbb{X}}(x, x') + 22r/3 + 4\varepsilon \geq d_{\mathbb{X}}(x, x')$. Putting all the above inequalities together we finally get

$$1 - \kappa(e) \leq \frac{l(\hat{e})}{l(e)} \leq 1 + \kappa'(e)$$

$$\text{with } \kappa(e) = \frac{10r}{3l(e)} + \left(\frac{5}{l(e)} + \frac{2}{R}\right)\varepsilon, \kappa'(e) = \left(\frac{3}{l(e)} + \frac{2}{R}\right)\varepsilon$$

Using that $l(e) \geq b$, we obtain that $\kappa(e) \leq \kappa = \frac{10r}{3b} + (\frac{5}{b} + \frac{2}{R})\varepsilon$ and $\kappa'(e) \leq (\frac{3}{b} + \frac{2}{R})\varepsilon$. As a consequence, since e and $\hat{e}$ are isometric to intervals, there exists a homeomorphism $\phi_e : e \to \hat{e}$ such that ϕ_e is $(1+\kappa')$-Lipschitz and ϕ^{-1} is $(1-\kappa)^{-1}$-Lipschitz. Since $\mathbb{X}$ and $\hat{\mathbb{X}}$ are graphs, the homeomorphisms ϕ_e can be glued all together to obtain a global homeomorphism $\phi : \mathbb{X} \to \hat{\mathbb{X}}$ such that ϕ is $(1+\kappa')$-Lipschitz and ϕ^{-1} is $(1-\kappa)^{-1}$-Lipschitz. □

Recall that to each edge $\hat{e}$ of $\hat{\mathbb{X}}$ we assign a length equal to the diameter of the corresponding connected component in $\mathcal{R}_{2r}(\mathbb{E})$ plus $4r$ and we denote by $d_{\hat{\mathbb{X}}}$ the metric induced on $\hat{\mathbb{X}}$. To conclude the metric reconstruction part, we finally relate the metrics on $\mathbb{Y}$ and $\hat{\mathbb{X}}$.

THEOREM 4. *There exists a map $\psi : \mathbb{Y} \to \hat{\mathbb{X}}$ such that for any $y, y' \in \mathbb{Y}$*

$$(1-\kappa)\left((1 - \frac{2\varepsilon}{R}) d_{\mathbb{Y}}(y, y') - \varepsilon\right) \leq d_{\hat{\mathbb{X}}}(\psi(y), \psi(y'))$$

$$\leq (1+\kappa')\left((1 + \frac{2\varepsilon}{R}) d_{\mathbb{Y}}(y, y') + \varepsilon\right)$$

with κ and κ' as in the Metric Reconstruction Theorem 2.

PROOF. Let C be an (ε, R)-correspondence between $\mathbb{Y}$ and $\mathbb{X}$. From the definition of correspondence, there exists a map (not necessarily continuous) $f : \mathbb{Y} \to \mathbb{X}$ such that for any $y \in \mathbb{Y}$, $(f(y), y) \in C$. It immediately follows from Lemma 1 and Theorem 2 that $\psi = \phi \circ f$ verifies the desired inequalities. □

Although the above result does not provide an explicit map, we provide an easy to compute map $\psi : \mathbb{Y} \to \hat{\mathbb{X}}$ that satisfies similar inequalities when restricted to edge components. First we define ψ on the branch points: each branch point is mapped to the vertex of $\hat{\mathbb{X}}$ corresponding to the connected component of $\mathcal{R}_{2r}(\mathbb{V})$ that contains it. We then define ψ on each connected component of $\mathcal{R}_{2r}(\mathbb{E})$. Let $\hat{e}$ be an edge of $\hat{\mathbb{X}}$ and let y_0, y_1 be two points in the corresponding connected component in $\mathcal{R}_{2r}(\mathbb{E})$ such that $d_{\mathbb{Y}}(y_0, y_1)$ is equal to the diameter of this component. We parametrize isometrically $\hat{e}$ by the interval $[0, l(\hat{e})]$. Recall that $l(\hat{e}) = d_{\mathbb{Y}}(y_0, y_1) + 4r$, we let $\psi(y_0) = 2r$ and $\psi(y_1) = l(\hat{e}) - 2r$. Now if $y \in \mathbb{Y}$ is in the same connected component of $\mathcal{R}_{2r}(\mathbb{E})$ as y_0 and y_1 we define $\psi(y) = 2r + d_{\mathbb{Y}}(y, y_0)\frac{d_{\mathbb{Y}}(y_0, y_1)}{d_{\mathbb{Y}}(y, y_0) + d_{\mathbb{Y}}(y, y_1)}$.

LEMMA 4. *For $i = 0, 1$ $(1 - \varepsilon M) d_{\mathbb{Y}}(y, y_i) \leq \psi(y) - \psi(y_i) \leq d_{\mathbb{Y}}(y, y_i)$ where $M = 6/R + 1/b$.*

PROOF. The proof of the case $i = 0$ and $i = 1$ being similar we give it for $i = 0$. Remark that $\psi(y) - \psi(y_0) = d_{\mathbb{Y}}(y, y_0)\frac{d_{\mathbb{Y}}(y_0, y_1)}{d_{\mathbb{Y}}(y, y_0) + d_{\mathbb{Y}}(y, y_1)}$ and the second inequality is just the triangle inequality. Let $x, x_0, x_1 \in \mathbb{X}$ be such that $(x, y), (x_0, y_0), (x_1, y_1) \in C$. Note that x, x_0 and x_1 are in the same edge of $\mathbb{X}$ so that $d_{\mathbb{X}}(x_0, x_1)$ can be expressed either as a sum or as a difference of $d_{\mathbb{X}}(x_0, x)$ and $d_{\mathbb{X}}(x, x_1)$. Applying Lemma 1 three times and using that $d_{\mathbb{Y}}(y_0, y) \leq d_{\mathbb{Y}}(y_0, y_1)$ and $d_{\mathbb{Y}}(y, y_1) \leq d_{\mathbb{Y}}(y_0, y_1)$ we obtain

$$d_{\mathbb{Y}}(y, y_0) + d_{\mathbb{Y}}(y, y_1) \leq d_{\mathbb{Y}}(y_0, y_1)\left(1 + \frac{6\varepsilon}{R}\right) + \varepsilon$$

Using that $b \leq d_{\mathbb{Y}}(y_0, y_1)$ we finally get $\frac{d_{\mathbb{Y}}(y, y_0) + d_{\mathbb{Y}}(y, y_1)}{d_{\mathbb{Y}}(y_0, y_1)} \leq 1 + (\frac{6}{R} + \frac{1}{b})\varepsilon$. □

From Lemma 4, we easily get the following corollary controlling the distortion on the metric induced by the restriction of ψ to the vertices of a connected component of $\mathcal{R}_{2r}(\mathbb{E})$.

COROLLARY 2. *If y, y' are in the same connected component of $\mathcal{R}_{2r}(\mathbb{E})$ corresponding to an edge $\hat{e}$ in $\hat{\mathbb{X}}$ then $d_{\mathbb{Y}}(y, y') - \varepsilon M l(\hat{e}) \leq \psi(y) - \psi(y') \leq d_{\mathbb{Y}}(y, y') + \varepsilon M l(\hat{e})$.*

6. EXPERIMENTS

We implemented our algorithm in C++ using the Boost Graph Library [18]. Experiments were conducted on a 2.33GHz Macbook Pro with 3GB of RAM. To assess the generality of our algorithm, we used four very different real world data sets: earthquake data, GPS traces, astronomical data and Image Webs. Table 1 summarizes our results, and a detailed discussion follows in this section.

	Earthquake	GPS Traces	Astronomical	Image Webs
Number of Original Vertices	1600	28434	9276	530
Number of Reconstructed Vertices	18	497	3651	112
Number of Original Edges	3983	41669	34890	1711
Number of Reconstructed Edges	9	5402	14808	409
Graph Reconstruction Time	5.2846	43.2249	14.0829	0.729667
Original Dist Comp Time	0.016386	0.777398	0.60322	0.021817
Approx Dist Comp Time	0.004696	0.029821	0.29148	0.013379
Dist Comp Time Speedup	249%	2507%	107%	63%
Mean Distortion	6.4%	2.4%	22%	27%
Median Distortion	8.8%	2.0%	19%	17%

Table 1: Our algorithm was used on several data sets to reconstruct a simpler metric graph approximating the distances in the original graph. We randomly selected a sample of 100 points and computed all pairwise distances between points in the same connected components. The graph computation time is the total time of estimating degrees of nodes and reconstructing the graph. The original computation time shows the total time of computing these distances using the original graph. The approximate computation time is the total time it took to compute approximate distances with the help of the reconstructed graph. All times are in seconds.

Data Sets

We used four different data sets for which we expect there to be an underlying metric graph approximation. The first data set is that of earthquake locations through which we wish to learn topological and geometric information about earthquake faults. The raw data was obtained from USGS Earthquake Search [1] and consists of earthquakes between 01/01/1970 and 01/01/2010, of magnitude greater than 5.0, and of location in the rectangular area between latitudes -75 degrees and 75 degrees and longitude between -170 degrees and 10 degrees. The underlying metric graph for this data set is the network of fault lines. The second data set is that of 500 GPS traces tagged "Moscow" from OpenStreetMap [2]. Since cars move on roads, we expect the locations of cars to provide information about the metric graph structure of the Moscow road network. The third data set consists of locations of galaxies in a portion of 3D space and there have been recent studies on the existence of filamentary structure in the distribution of galaxies [9]. Lastly, we include an *Image Web* [15] data set which is a collection of images, with similar regions linked together to form a graph structure. Dense image collections are often acquired by mobile entities, and thus naturally contain long linear and circular parts, joined together at branch points.

Preprocessing and Parameter Selection

We performed some preprocessing to transform the raw data into a metric space $(\mathbb{Y}, d_{\mathbb{Y}})$ on which we could use our algorithm to discover a much simpler metric graph $(\mathbb{X}, d_{\mathbb{X}})$ approximating this space. Since real world data sets vary widely in both noise and scale, the specific preprocessing steps differ across the data sets. However, in most of our examples, we first construct a neighborhood graph on the data, and then used the shortest path metric space on the neighborhood graph as the input to our algorithm. The raw earthquake data set contains the coordinates of the epicenters of 12790 earthquakes in the latitude/longitude rectangle $[-75, 75] \times [-170, 10]$. As it contains outliers, we first preprocessed the data by removing points located in low density areas using the *distance-to-measure function* [7]: points with average squared distance to their 30 nearest neighbors larger than $4.7^2 = 22.09$ were discarded, resulting in the elimination of 284 points. Among the remaining data, points with average squared distance to their 50 nearest neighbors larger than 81 were also discarded to get a cleaner data set (eliminating 41 more points). Then, we randomly sampled 1600 landmarks among the points with average squared distance to their 50 nearest neighbors in the cleaned data set smaller than 1.5. Finally, we computed an α-complex with $\alpha = 4$ on these landmarks, and used the shortest path metric on this complex as the input to our algorithm. For the road network data set, we first selected a metric ε-net on the raw GPS locations with $\varepsilon = 5$ using furthest point sampling. Then, we computed an α-complex on the ε-net as the neighborhood graph, but with $\alpha = 50$. The astronomical data is similar to the earthquake data in the sense that it contains a lot of noise, which hides the filament structure. We built the input neighborhood graph on a set of landmarks selected in a similar fashion as that for the earthquake data set. The Image Web data set differs from the rest in that the raw data is a neighborhood graph, so no preprocessing was done.

Our algorithm is parameterized by the spherical shell inner radius r, which in the analysis is allowed to be in a range of values that depends on a constant b that is the property of the underlying metric graph and the level of approximation attained by the data. In practice, however, we do not have an oracle for these constants. However, regardless of whether the assumptions in the analysis are satisfied, our implementation outputs a metric graph $(\hat{\mathbb{X}}, d_{\hat{\mathbb{X}}})$ and a map ϕ from the raw data to the metric graph. Using random sampling, we can estimate the level of metric distortion using $\hat{\mathbb{X}}$ and ϕ. Thus, we are able to select the parameter by running our algorithm using various values of r, and checking for a balance between metric distortion and reduction of graph size. We note that even though the assumptions in the analysis may not be strictly satisfied, our algorithm returns a metric graph approximation that is in some cases dramatically smaller than the original data, while approximately preserving distances. In addition, we also varied the outer radius ($5r/3$ in the analysis) and the Rips-Vietoris parameter ($4r/3$ in the analysis) using the same process. Indeed, the constants $4r/3$ and $5r/3$ were chosen for ease of analysis. In particular, they ensured than in the proof of Theorem 3,

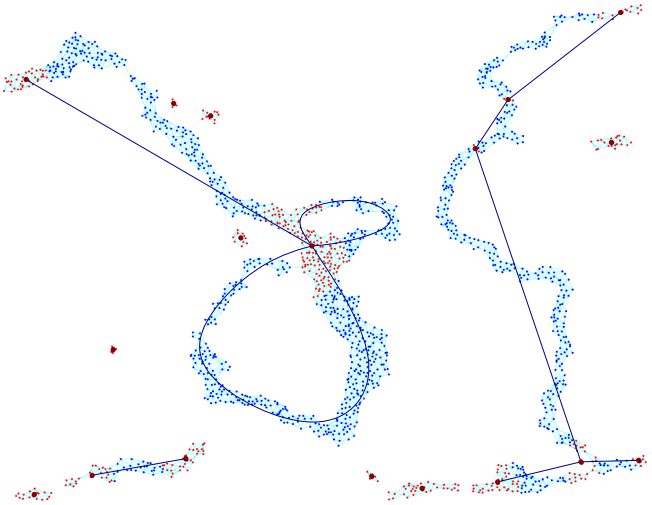

Figure 2: Earthquake data: the input neighborhood graph is shown in cyan, the points marked as belonging to a branch are shown in red, and the points marked as belonging to an edge are shown in blue. The reconstructed graph is shown in dark blue.

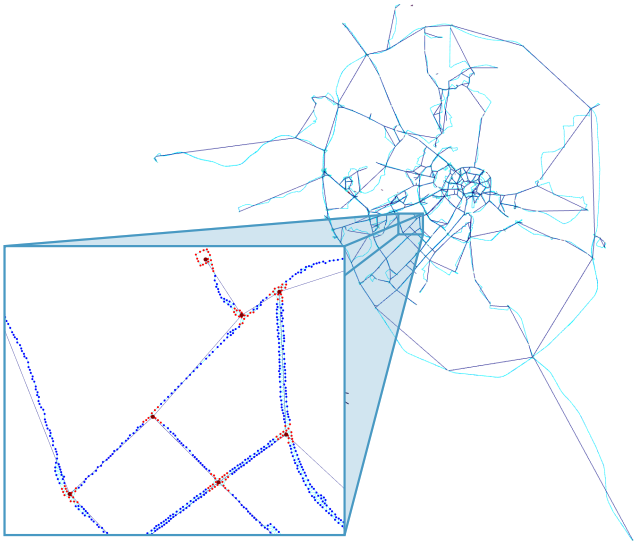

Figure 3: GPS traces: the input neighborhood graph is shown in cyan, the points marked as belonging to a branch are shown in red, and the points marked as belonging to an edge are shown in blue. The reconstructed graph is shown in dark blue.

all connected components were cliques, but in practice they may not be the best constants to use.

Implementation and Results

Real world data sets often do not satisfy the assumptions we require for complete reconstruction, so we only replace connected components of $\mathcal{R}_{2r}(\mathbb{E})$ with edges of $\hat{\mathbb{X}}$ if they are adjacent to exactly one (in the case of a self-loop) or two connected components of $\mathcal{R}_{2r}(\mathbb{V})$. Note that this process is local and hence it is possible to iterate this process in order to discover stratified structure at multiple scales. We also computed a map ψ from the original points to the reconstructed space $\hat{\mathbb{X}}$ as described in Lemma 4. To evaluate the quality of the reconstructed graph for each data set, we randomly selected 100 points from the data set, and computed both original pairwise distances, and pairwise distances on $\hat{\mathbb{X}}$ using ψ. We also evaluated the use of $\hat{\mathbb{X}}$ to speed up distance computations by showing reductions in computation time. Statistics for the size of the reconstructed graph, error of approximate distances, and reduction in computation time are given in Table 1. Only pairs of vertices in the same connected component are included because we obtain zero error for the pairs of vertices that are not. We used these statistics to select the parameter r, as well as the outer radius of the spherical shell, and the Rips-Vietoris parameter.

The result of our algorithm on the earthquake data set is shown in Fig. 2. We observe two spurious branch points being detected on the component to the right as a result of the small stub sticking out between them. Nevertheless, our algorithm is able to replace the data by a much smaller graph, while maintaining small distortion of distance. Note that a trivial postprocessing step that removes all vertices of degree 2 could take care of the two spurious branch points. The GPS trace data set, shown in Fig. 3, provides the best results of all four data sets, showing a dramatic reduction in graph size along with a very small distortion of distance. This is expected considering that cars in most cities necessarily follow a road network, which fits the model of a metric graph very well. The metric graph structure in the astronomical data set, shown in Fig. 4, is much less apparent than that of the previous examples, and hence we were only able to reduce the graph size by one half. However, by doing so, we still approximately preserved distances and reduced distance computation time by more than 51%. The Image Web, shown in Fig. 5, was a very small example, and therefore suffers from metric distortion problems as noise levels are relatively large when compared to the size of the branching structures, but our algorithm was still able to reduce the already small graph size by 79% while keeping the median distance distortion below 18%.

7. CONCLUSION AND FUTURE WORK

In this paper, we presented a first attempt at reconstructing a metric space of mixed dimension. We presented an algorithm with guarantees for the case of a metric graph, or equivalently, a 1-D stratified space. The same algorithm can be used to simplify the representation of a metric space that might possibly have an underlying metric graph structure. We also showed that, on real world data that doesn't perfectly satisfy the hypotheses, our algorithm still gives sensible and useful results.

A natural extension of this work would be to consider stratified spaces of higher dimension, as well as considering the data at multiple scales. Currently, we rely on the fact that our algorithm is relatively fast, and thus trying various scale parameters and checking for a small reconstructed metric graph with small distortion is feasible. However, it is also interesting to consider the automatic selection of scales for which the data can be viewed as a reasonable approximation of a metric graph. We have also begun preliminary experiments for a multiscale version of our algorithm, which follows naturally from our implementation. It would be of interest

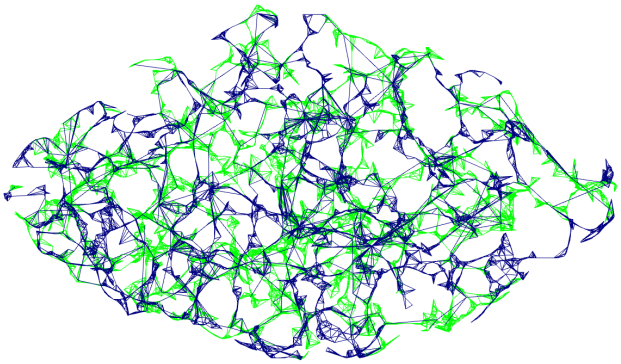

Figure 4: Astronomical data: the input neighborhood graph is shown in green and the reconstructed graph is shown in dark blue.

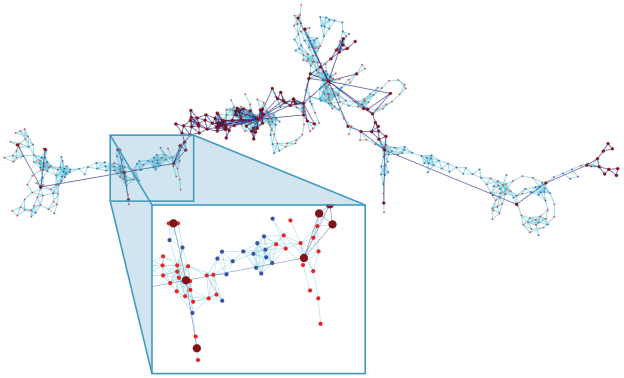

Figure 5: Image Web: the input neighborhood graph is shown in cyan, the points marked as belonging to a branch are shown in red, and the points marked as belonging to an edge are shown in blue. The reconstructed graph is shown in dark blue.

to consider models of data where such a reconstruction gives theoretical guarantees. Other directions for further research include investigating the possibility of improving the distortion of the metric by allowing the addition of branch points to split edges that have too much distortion or to contract large regions of branch points into several points instead of just one. Having these options not only gives the user some choice on the trade-off between the size of the graph and the distortion, but also fits well with a multiscale approach.

8. ACKNOWLEDGMENTS

This work was supported in part by NSF grants CCF 0634803, FODAVA 0808515, CCF 1011228, a grant from Google Inc., EU project CGLearning 255827, and ANR grant GIGA ANR-09-BLAN-0331-01. We would also like to thank the associated research team CoMeT.

9. REFERENCES

[1] Earthquake search. http://earthquake.usgs.gov/earthquakes/eqarchives/epic/.

[2] Openstreetmap. http://www.openstreetmap.org/.

[3] H. Abramowicz, D. Horn, U. Naftaly, and C. Sahar-Pikielny. Orientation selective neural network for cosmic muon identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2):163–166, 1997. New Computing Techniques in Physics Research V.

[4] N. Amenta, M. Bern, and D. Eppstein. The crust and the β-skeleton: Combinatorial curve reconstruction. *Graph. Models Image Process.*, 60(2):125–135, 1998.

[5] E. Arias-Castro, D. L. Donoho, , and X. Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Annals of Statistics*, 34(1):326–349, 2006.

[6] D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[7] F. Chazal, D. Cohen Steiner, and Q. Mérigot. Geometric Inference for Measures based on Distance Functions. Research Report RR-6930, INRIA, 2010.

[8] D. Chen, L. J. Guibas, J. Hershberger, and J. Sun. Road network reconstruction for organizing paths. In *Proceedings 21st ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

[9] E. Choi, N. A. Bond, M. A. Strauss, A. L. Coil, M. Davis, and C. N. A. Willmer. Tracing the filamentary structure of the galaxy distribution at $z \sim 0.8$. *Monthly Notices of the Royal Astronomical Society*, pages 692–+, May 2010.

[10] T. K. Dey, K. Mehlhorn, and E. A. Ramos. Curve reconstruction: Connecting dots with good reason. In *Proceedings of the Fifteenth Annual Symposium on Computational Geometry*, pages 197–206. ACM, 1999.

[11] T. K. Dey and R. Wenger. Reconstructing curves with sharp corners. *Comput. Geom. Theory Appl.*, 19:89–99, July 2001.

[12] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. On the path density of a gradient field. *Annals of Statistics*, 37(6A):3236–3271, 2009.

[13] C. R. Genovese, M. Perone-Pacifico, I. verdinelli, and L. Wasserman. Nonparametric Filament Estimation. *ArXiv e-prints*, Mar. 2010.

[14] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser, 2nd edition, 2007.

[15] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. J. Guibas. Image webs: Computing and exploiting connectivity in image collections. *Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[16] P. Kuchment. Quantum graphs I. Some basic structures. *Waves in Random Media*, 14(1):S107–S128, 2004.

[17] N. Qaddoumi, E. Ranu, J. D. McColskey, R. Mirshahi, and R. Zoughi. Microwave detection of stress-induced fatigue cracks in steel and potential for crack opening

determination. *Research in Nondestructive Evaluation*, 12(2):87–104, 2000.

[18] J. Siek, L.-Q. Lee, and A. Lumsdaine. Boost graph library. http://www.boost.org/libs/graph/, June 2000.

[19] F. Tupin, H. Maitre, Mangin, N. J.-F., J.-M., and E. Pechersky. Detection of linear features in SAR images: Application to road network extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 36:434ö–453, 1998.

APPENDIX

A. (R,α)-DEGREE INFERENCE

As mentioned in Section 5.1, we can make the r-degree dependent of an extra parameter $0 < \alpha < 1$ and prove a similar result as Theorem 3.

Definition 5. Let $(\mathbb{Y}, d_{\mathbb{Y}})$ be an (ε, R)-approximation of X. Given $0 < r < R/2$ and $0 \leq \alpha < 1$, the (r,α)-degree $deg_{r,\alpha}(y)$ of a point $y \in \mathbb{Y}$ is the number of connected components of the Rips-Vietoris graph with parameter $2\alpha r$ and vertex set $\mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ where $\mathrm{B}_{\mathbb{Y}}(y,r)$ denotes the intrinsic (closed) ball in Y with center y and radius r.

Theorem 5 (Degree Inference Theorem). *Let $(\mathbb{Y}, d_{\mathbb{Y}})$ be an (ε, R)-approximation of $\mathbb{X}$. Let $C \subset \mathbb{X} \times \mathbb{Y}$ be an (ε, R)-correspondence between $\mathbb{X}$ and Y, let $(x,y) \in C$ and let $0 < \alpha < 1$.*
i) If the distance d_0 from x to any vertex of $\mathbb{X}$ is larger than $(3\max(\frac{1+\alpha}{\alpha}, \frac{1+\alpha}{2(1-\alpha)}) + 1)\varepsilon$ then for $3\max(\frac{1}{\alpha}, \frac{1}{2(1-\alpha)})\varepsilon < r < \min(\frac{R}{2}, \frac{d_0-\varepsilon}{1+\alpha})$, $deg_{r,\alpha}(y)$ is equal to the degree of x in $\mathbb{X}$ (i.e. 2). Moreover the pairwise distances between the connected components of the Rips-Vietoris graph are lower bounded by $2r - 3\varepsilon$.
ii) If x is at distance less than ε from a vertex x_0 of $\mathbb{X}$ and if the length l_0 of the shortest edge adjacent to x_0 is larger than $(\max(\frac{3(1+\alpha)}{\alpha}, \frac{5(1+\alpha)}{2(1-\alpha)}) + 1)\varepsilon$ then for $\max(\frac{3}{\alpha}, \frac{5}{2(1-\alpha)})\varepsilon < r < \min(\frac{R}{2}, \frac{l_0-2\varepsilon}{1+\alpha})$, $d_{r,\alpha}(y)$ is equal to the degree of x_0 in $\mathbb{X}$. Moreover the pairwise distances between the connected components of the Rips-Vietoris graph are lower bounded by $2r - 5\varepsilon$.

This result motivates the choice of the value $\alpha = 2/3$ in the paper: this is the value that minimizing the bound $(3\max(\frac{1+\alpha}{\alpha}, \frac{1+\alpha}{2(1-\alpha)}) + 1)\varepsilon$ in *i)* that controls the size of the expansion procedure in Section 5.2.

Proof. The proof of the above theorem being almost verbatim the same as the one of Theorem 3, it is just given for completeness.
First remark that if $(x',y') \in C$ is such that $y' \in \mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ then $x' \in \mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$.
i) Since $r > \varepsilon$ and $(1+\alpha)r + \varepsilon < d_0$, $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$ is included in the edge containing x and has exactly 2 connected components. Moreover, these two connected components are at distance $2(r-\varepsilon)$.

Now, if $(x',y'),(x'',y'') \in C$ are such that $y', y'' \in \mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ and $d_{\mathbb{Y}}(y',y'') < 2\alpha r$ then $d_{\mathbb{X}}(x',x'') < 2\alpha r + \varepsilon$ and, since $r > \frac{3\varepsilon}{2(1-\alpha)}$, it follows that x' and x'' are in the same connected component of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$.

Reciprocally, if $(x',y'),(x'',y'') \in C$ are such that x', x'' are in the same connected component of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$, then $d_{\mathbb{X}}(x',x'') \leq \alpha r + 2\varepsilon$ and $d_{\mathbb{Y}}(y',y'') \leq \alpha r + 3\varepsilon < 2\alpha r$ since $\alpha r > 3\varepsilon$.

As a consequence, the Rips-Vietoris graph with parameter $2\alpha r$ and vertex set $\mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ has at most two connected components. To prove that it has exactly two connected components one just needs to check that each connected component K of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$ contains a point x' such that there exists $y' \in \mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ satisfying $(x',y') \in C$: let x' be the point of K such that $d_{\mathbb{X}}(x,x') = (1+\alpha/2)r$ and let $(x',y') \in C$. Then, since $\alpha r > 2\varepsilon$, $d_{\mathbb{Y}}(y,y') \leq (1+\alpha/2)r + \varepsilon < (1+\alpha)r$ and $d_{\mathbb{Y}}(y,y') \geq (1+\alpha/2)r - \varepsilon > r$.
ii) This is almost the same proof as for *i)* except that since x is not a vertex, but at distance at most ε from a vertex we have to slightly change the constraint on r.

Let $y_0 \in \mathbb{Y}$ be such that $(x_0,y_0) \in C$. From the definition of $(\varepsilon,0)$-approximation we have $d_{\mathbb{Y}}(y,y_0) < 2\varepsilon$.

Since $r > 2\varepsilon$ and $(1+\alpha)r + 2\varepsilon < l_0$, $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$ has exactly d connected components, each included in different edges adjacent to x_0, where d is the degree of x_0. Moreover these connected components are at distance a least $2(r-2\varepsilon)$ from each other.

Now, if $(x',y'),(x'',y'') \in C$ are such that $y', y'' \in \mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ and $d_{\mathbb{Y}}(y',y'') < 2\alpha r$ then $d_{\mathbb{X}}(x',x'') < 2\alpha r + \varepsilon$ and, since $r > \frac{5\varepsilon}{2(1-\alpha)}$, it follows from claim 1 that x' and x'' are in the same connected component of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$.

Reciprocally, if $(x',y'),(x'',y'') \in C$ are such that x', x'' are in the same connected component of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$, then $d_{\mathbb{X}}(x',x'') \leq \alpha r + 2\varepsilon$ and $d_{\mathbb{Y}}(y',y'') \leq \alpha r + 3\varepsilon < 2\alpha r$ since $\alpha r > 3\varepsilon$.

As a consequence, the Rips-Vietoris graph with parameter $2\alpha r$ and vertex set $\mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ has at most d connected components. To prove that it has exactly d connected components one just needs to check that each connected component K of $\mathrm{B}_{\mathbb{X}}(x,(1+\alpha)r+\varepsilon) \setminus \mathrm{B}_{\mathbb{X}}(x,r-\varepsilon)$ contains a point x' such that there exists $y' \in \mathrm{B}_{\mathbb{Y}}(y,(1+\alpha)r) \setminus \mathrm{B}_{\mathbb{Y}}(y,r)$ satisfying $(x',y') \in C$: let x' be the point of K such that $d_{\mathbb{X}}(x,x') = (1+\alpha/2)r$ and let $(x',y') \in C$. Then, since $\alpha r > 2\varepsilon$, $d_{\mathbb{Y}}(y,y') \leq (1+\alpha/2)r + \varepsilon < (1+\alpha)r$ and $d_{\mathbb{Y}}(y,y') \geq (1+\alpha/2)r - \varepsilon > r$. □