Metric denoising: Making it more friendly for topological computation

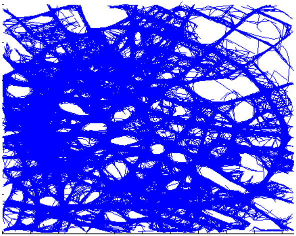
Yusu Wang

The Ohio State University

TDA-BIRS 2017

Introduction

- Noise in data prevalent in various applications
- Noise present in diverse forms
- Effective handling of noise depends on how they are generated and what the target uses of data are



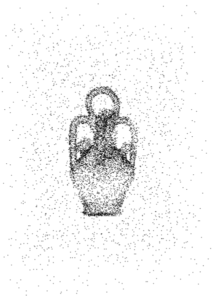
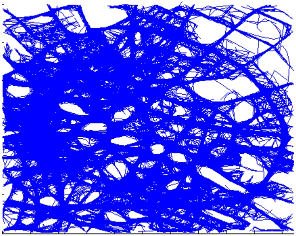


Image from *brainmaps.org*

Introduction

- Noise in data prevalent in various applications
- Noise present in diverse forms
- Effective handling of noise depends on how they are generated and what the target uses of data are
- This talk:
  - Focus on noise in metric of input data



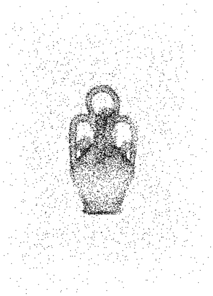


Image from *brainmaps.org*

In Topological Data Analysis

- Many geometric / topological data analysis algorithms often assume that the input is a finite metric space.
  - One of the most popular setting: input is point clouds data embedding in an ambient (Euclidean) space

In Topological Data Analysis

- Many geometric / topological data analysis algorithms often assume that the input is a finite metric space.
  - One of the most popular setting: input is point clouds data embedding in an ambient (Euclidean) space
- Limited types of noise that can be handled:
  - Mostly (Gromov)-Hausdorff type noise

In Topological Data Analysis

- Many geometric / topological data analysis algorithms often assume that the input is a finite metric space.
  - One of the most popular setting: input is point clouds data embedding in an ambient (Euclidean) space
- Limited types of noise that can be handled:
  - Mostly (Gromov)-Hausdorff type noise

Theorem (An exapmle)

Given two sets of points $P, Q \subseteq \mathbb{R}^d$, let dgm P and dgm Q denote the persistence diagrams induced by the Čech filtration on P and Q, respectively. Then

$$d_B(dgm\ P, dgm\ Q) \leq d_H(P, Q).$$

In Topological Data Analysis

- Many geometric / topological data analysis algorithms often assume that the input is a finite metric space.
    - One of the most popular setting: input is point clouds data embedding in an ambient (Euclidean) space
- Limited types of noise that can be handled:
    - Mostly (Gromov)-Hausdorff type noise
    - Some work handle more general noise, e.g, work on distance to measures [Chazal, Cohen-Steiner, Mérigot, 2011]

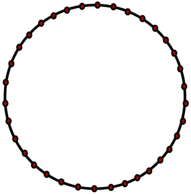
In Topological Data Analysis

- Many geometric / topological data analysis algorithms often assume that the input is a finite metric space.
  - One of the most popular setting: input is point clouds data embedding in an ambient (Euclidean) space
- Limited types of noise that can be handled:
  - Mostly (Gromov)-Hausdorff type noise
  - Some work handle more general noise, e.g, work on distance to measures [Chazal, Cohen-Steiner, Mérigot, 2011]
- Averaging in the space of persistence diagrams may not be effective.

A graph exampe

Suppose our input are observed graphs

- Say, graphs G_1, $G_2, \ldots,$ are noisy observation of the same true graph G^*
- We may try to build intrinsic Čech filtration based on induced graph metric and then "average" their persistence diagrams $dgm_1, dgm_2, \ldots$

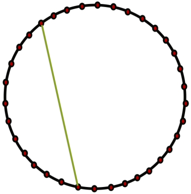


G^*

A graph exampe

Suppose our input are observed graphs

- Say, graphs G_1, $G_2, \ldots,$ are noisy observation of the same true graph G^*
- We may try to build intrinsic Čech filtration based on induced graph metric and then "average" their persistence diagrams $dgm_1, dgm_2, \ldots$

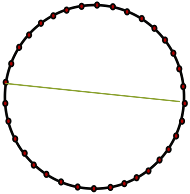


G_1

A graph exampe

Suppose our input are observed graphs

- Say, graphs G_1, $G_2, \ldots,$ are noisy observation of the same true graph G^*
- We may try to build intrinsic Čech filtration based on induced graph metric and then "average" their persistence diagrams $dgm_1, dgm_2, \ldots$

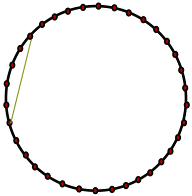


G_2

A graph exampe

Suppose our input are observed graphs

- Say, graphs G_1, $G_2, \ldots,$ are noisy observation of the same true graph G^*
- We may try to build intrinsic Čech filtration based on induced graph metric and then "average" their persistence diagrams $dgm_1, dgm_2, \ldots$



G_3

Our goal

To facilitate TDA tasks, our goal is to

- denoise input metric so that it is close to the “true” metric under Hausdorff-type distance

Overview

Three different settings to explore:

What are natural ways to model noise in input metric, and how to process such noise effeciently with theoretical guarantees.

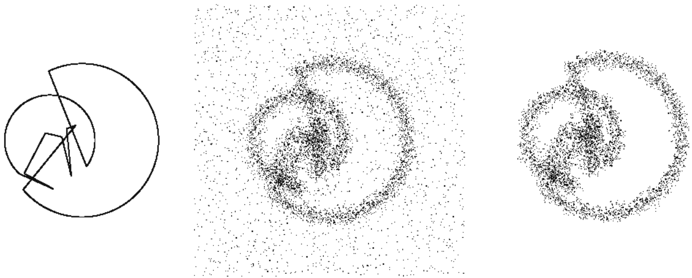
- Setting 1: towards parameter-free denoising for embedded point cloud data (PCD)
- Setting 2: metric embedding with outliers
- Setting 3: recovering shortest path metrics from perturbed graphs

Setting 1

Input: A set of points P *already embedded* in a metric space, which is a *"noisy"* sample of a hidden ground truth K

Output: A *"denoised"* set of points $Q \subset P$ Hausdorff-close to K

- [Buchet, Dey, J. Wang, W. SoCG 2017]

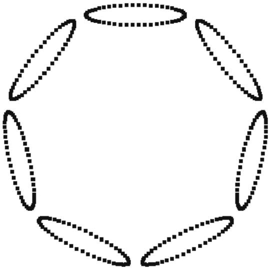


Some Existing Denoising Approaches

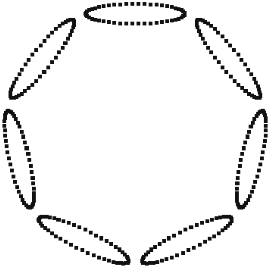
- Thresholding
  - choice of a density estimator, which involves parameter(s)
  - choice of a threshold
- Mean-shift type
  - needs additional parameters: such as step size, stopping criteria.
- Parametric methods
  - assuming knowing the noise distribution or generative model
  - often asymptotic guarantees

Require parameters and / or knowledge of noise models.
Non-uniform distribution challenging.

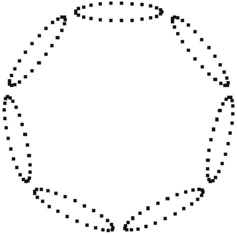
Parameter/assumptions Necessary



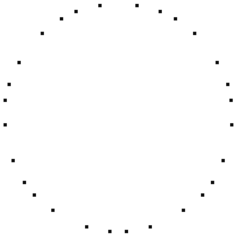
Parameter/assumptions Necessary



$k = 2$



$k = 10$



Goal of Setting-1

Minimize the use of parameter in denoising embedded PCD data, yet still provide theoretical guarantees / understanding

Goal of Setting-1

Minimize the use of parameter in denoising embedded PCD data, yet still provide theoretical guarantees / understanding

- Decluttering algorithm (works for any input, use one parameter)

Goal of Setting-1

Minimize the use of parameter in denoising embedded PCD data, yet still provide theoretical guarantees / understanding

- Decluttering algorithm (works for any input, use one parameter)

Parameter-free? Require stronger assumptions on noise model

Goal of Setting-1

Minimize the use of parameter in denoising embedded PCD data, yet still provide theoretical guarantees / understanding

- Decluttering algorithm (works for any input, use one parameter)

Parameter-free? Require stronger assumptions on noise model

- Declutter+Resample algorithm

Theoretical guarantees for decluttering

Theorem

Given a point set P which is an ϵ_k noisy sample of a compact K, Algorithm Declutter *returns a set Q such that*

$$d_H(K, Q) \leq 7\epsilon_k.$$

Theoretical guarantees for decluttering

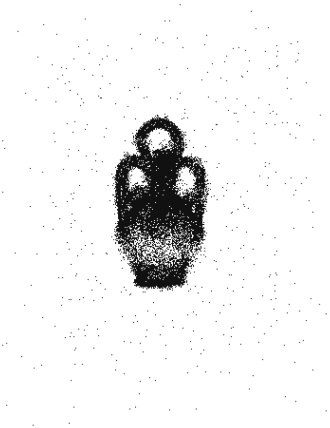
Theorem

Given a point set P which is an ϵ_k noisy sample of a compact K, Algorithm Declutter *returns a set Q such that*

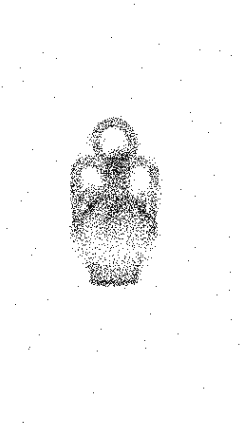
$$d_H(K, Q) \leq 7\epsilon_k.$$

- Can be extended to an *adaptive-noise* setting

Illustration II



Input



$k = 4$

$k = 47$

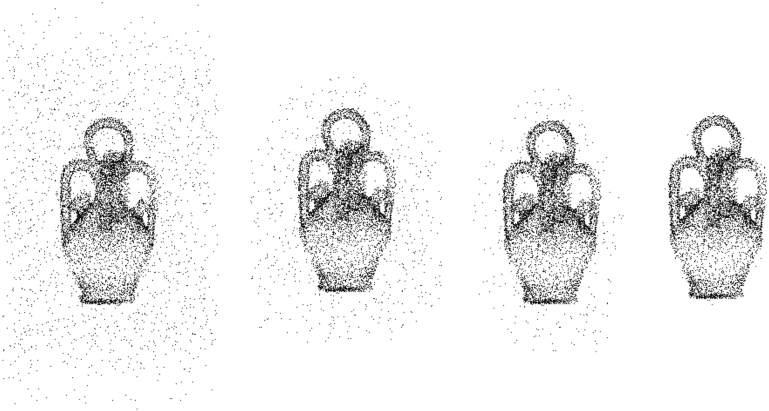
ParaFreeDecluster guarantees

Theorem

Given a point set P and i_0 such that for all $i > i_0$, P is a weak uniform $(\epsilon_{2^i}, 2)$ noisy sample of K and is also an $(\epsilon_{2^{i_0}}, 2)$ noisy sample of K, Algorithm ParfreeDeclutter *returns a point set P_0 such that $d_H(P_0, K) \leq (87 + 16\sqrt{2})\epsilon_{2^{i_0}}$.*

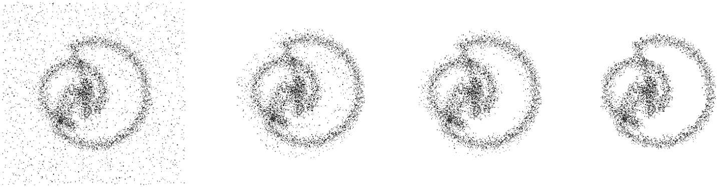
- Require uniformity of input samples around the hidden compact set.
- Algorithm still very simple. It has $O(\log n)$ iterations of previous Declutter algorithm and another resampling procedure.

ParaFreeDeclutter results



Input $k = 1024$ $k = 256$ $k = 1$

ParaFreeDeclutter results



Setting 2

Input: A discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$

- (X, ρ) approximately comes from a "nice" *target metric space*
- some input points could have corrupted / erroneous distance to other points, they are "outliers"

Setting 2

Input: A discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$

- (X, ρ) approximately comes from a "nice" *target metric space*
- some input points could have corrupted / erroneous distance to other points, they are "outliers"

Output: A "near-optimal" set of outliers $K \subset X$ together with a "low-distortion" embedding of $(X \setminus K, \rho)$ into some target metric space

- the target space could be a tree metric, ultrametric, or constant-dimensional Euclidean space.

[Sidiropoulos, D. Wang, W. SoDA 2017]

Notations

Definition (Embedding)

Given two metric spaces $\mathcal{X} = (X, \rho_X)$ and $\mathcal{Y} = (Y, \rho_Y)$, an *embedding* of $\mathcal{X}$ into $\mathcal{Y}$ is simply a map $\phi : X \to Y$.

- ϕ is an *isometric embedding* if for any $x, x' \in X$, $\rho_X(x, x') = \phi_Y(\phi(x), \phi(x'))$.
- ϕ is an *ε-distorted embedding* if for any $x, x' \in X$, $|\rho_X(x, x') - \rho_Y(\phi(x), \phi(x')| \leq \varepsilon$. Alternatively, we say that $\mathcal{X}$ admits an embedding into $\mathcal{Y}$ with (additive) distortion ε.

Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

$$
\begin{array}{c c}
 & \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\
\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} &
\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{bmatrix}
\end{array}
$$

Input metric $\mathcal{X}$

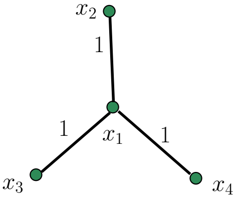
Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

	x_1	x_2	x_3	x_4
x_1	0	1	1	1
x_2	1	0	2	2
x_3	1	2	0	2
x_4	1	2	2	0

Input metric $\mathcal{X}$



Isometric embedding to a tree metric

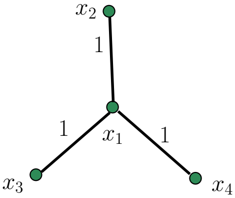
Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

	x_1	x_2	x_3	x_4
x_1	0	1.1	0.89	1.05
x_2	1.1	0	2.12	1.95
x_3	0.89	2.12	0	2
x_4	1.05	1.95	2	0

Input metric $\mathcal{X}'$



Isometric embedding to a tree metric

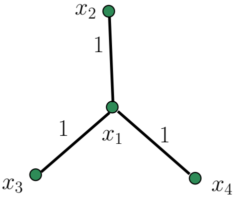
Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

	x_1	x_2	x_3	x_4
x_1	0	1.1	0.89	1.05
x_2	1.1	0	2.12	1.95
x_3	0.89	2.12	0	2
x_4	1.05	1.95	2	0

Input metric $\mathcal{X}'$



An embedding with additive distortion 0.2

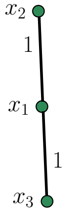
Optimization Problem

Minimum outlier-embedding problem: Given a discrete n-point metric space $(X = \{x_1, \ldots, x_n\}, \rho)$, compute the *smallest set* $K^* \subset X$ such that $(X \setminus K^*, \rho)$ embeds into a target metric space either isometrically, or with distortion at most ε.

- Choices of target metric spaces: ultrametric, tree metric, constant-dimensional Euclidean space $\mathbb{R}^d$
- The set K^* is refered to as the *optimal set of outliers*

	x_1	x_2	x_3	x_4
x_1	0	1.1	0.89	1.05
x_2	1.1	0	2.12	1.95
x_3	0.89	2.12	0	2
x_4	1.05	1.95	2	0

Input metric $\mathcal{X}'$



Outlier embedding to $\mathbb{R}^1$ with distortion 0.13

Hardness of the Outlier Embedding

Theorem

The problem of minimum outlier embedding into a tree metric, an ultrametric, or $\mathbb{R}^d$, is NP-hard.

Furthermore, assuming the Unique Games Conjecture, it is NP-hard to approximate the isometric version with a factor of $2 - \nu$ for any $\nu > 0$.

Hardness of the Outlier Embedding

Theorem

The problem of minimum outlier embedding into a tree metric, an ultrametric, or $\mathbb{R}^d$, is NP-hard.

Furthermore, assuming the Unique Games Conjecture, it is NP-hard to approximate the isometric version with a factor of $2 - \nu$ for any $\nu > 0$.

Our next goal

Efficient approximation algorithms for the outlier-embedding problems.

- We developed various approximation algorithms

Hardness of the Outlier Embedding

Theorem

The problem of minimum outlier embedding into a tree metric, an ultrametric, or $\mathbb{R}^d$, is NP-hard.

Furthermore, assuming the Unique Games Conjecture, it is NP-hard to approximate the isometric version with a factor of $2 - \nu$ for any $\nu > 0$.

Our next goal

Efficient approximation algorithms for the outlier-embedding problems.

- We developed various approximation algorithms
- Present results for special case: *(near-)isometric outlier-embedding into $\mathbb{R}^d$*

Isometric outlier-embedding case

Theorem (First 2-approximation)

Given an n-point metric space (X, ρ), there is an algorithm that can compute at most $2|K^*|$ number of points $K \subset X$, such that $(X \setminus K, \rho)$ admits an isometric embeddign into $\mathbb{R}^d$. The algorithm runs in $O(n^{d+1})$ time.

Isometric outlier-embedding case

Theorem (First 2-approximation)

Given an n-point metric space (X, ρ), there is an algorithm that can compute at most $2|K^*|$ number of points $K \subset X$, such that $(X \setminus K, \rho)$ admits an isometric embeddign into $\mathbb{R}^d$. The algorithm runs in $O(n^{d+1})$ time.

Theorem (Improved Approximation)

Given an n-point metric space (X, ρ), there is a $O(n^2)$ time randomized algorithm that can compute $3|K^*|$ number of points $K \subset X$, such that with constant probability, $(X \setminus K, \rho)$ admits an isometric embeddign into $\mathbb{R}^d$.

- The big O notation hides constants depending exponentially on the dimension d.

Low-distortion Case

Theorem (Bicriteria-Approximation)

Given an n-point metric space (X, ρ), suppose it admits an $X \setminus K^*$ admits a δ^*-distortion embedding into $\mathbb{R}^d$. Then there is a $O(n^2)$ time randomized algorithm that can compute $O(|K^*|d)$ number of points $K \subset X$, such that with constant probability, $(X \setminus K, \rho)$ admits an embeddign into $\mathbb{R}^d$ with distortion $O(\sqrt{\delta^*})$-distortion.

- The big O notation hides constants depending exponentially on the dimension d.

Low-distortion Case

Theorem (Bicriteria-Approximation)

Given an n-point metric space (X, ρ), suppose it admits an $X \setminus K^*$ admits a δ^*-distortion embedding into $\mathbb{R}^d$. Then there is a $O(n^2)$ time randomized algorithm that can compute $O(|K^*|d)$ number of points $K \subset X$, such that with constant probability, $(X \setminus K, \rho)$ admits an embeddign into $\mathbb{R}^d$ with distortion $O(\sqrt{\delta^*})$-distortion.

- The big O notation hides constants depending exponentially on the dimension d.
- Algorithm still reasonably simple, but analysis is much more involved.
  - We have implemented it!

Talk Outline

In this talk, we consider three different settings to explore:

What are natural ways to model noise in input metric, and how to process such noise effeciently with theoretical guarantees.

- Setting 1: towards parameter-free denoising for embedded point cloud data (PCD)
- Setting 2: metric embedding with outliers
- Setting 3: recovering shortest path metric from perturbed graphs

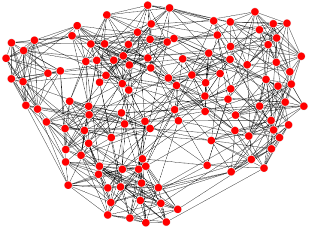
Problem Setup

Input: An observed unweighted graph $G = (V, E)$

- G is a "noisy" observation of a true graph G^*
- the metric of interest is the shortest path metric d_{G^*}

Output: Recover (approximately) the "true" shortest path metric d_{G^*} from G

- [Parthasarathy, Sivakoff, Tian, W. 2017]



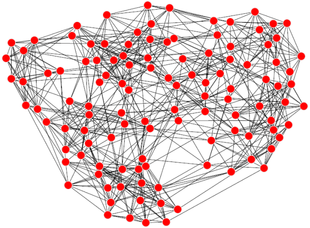
Problem Setup

Input: An observed unweighted graph $G = (V, E)$

- G is a *"noisy"* observation of a *true graph G^**
- the metric of interest is the shortest path metric d_{G^*}

Output: Recover (approximately) the "true" shortest path metric d_{G^*} from G

- [Parthasarathy, Sivakoff, Tian, W. 2017]



The model

The true graph $G^* = (V, E^*)$: given n,

- $V = V_n$ sampled i.i.d from a L-doubling measure $\mu : M \to \mathbb{R}^+$ on a compact geodesic metric space (M, d_M)
- $E^* = E^*_{r,n} = \{(u, v) \mid d_M(u, v) \leq r, u, v \in V\}$ is the r-neighborhood graph for some parameter $r > 0$

The model

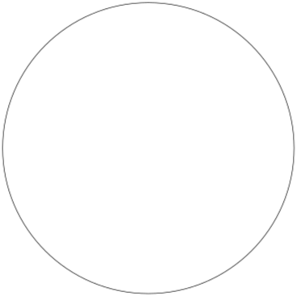
The true graph $G^* = (V, E^*)$: given n,

- $V = V_n$ sampled i.i.d from a L-doubling measure $\mu : M \to \mathbb{R}^+$ on a compact geodesic metric space (M, d_M)
- $E^* = E^*_{r,n} = \{(u, v) \mid d_M(u, v) \leq r, u, v \in V\}$ is the r-neighborhood graph for some parameter $r > 0$

The observed graph $G = (V, E)$: A (p, q)-perturbation of G^* where

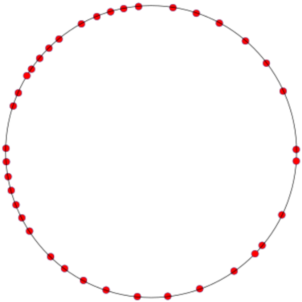
- (p-deletion): For each edge $e = (u, v) \in E^*$, we have $e \in E$ with probability $1 - p$
- (q-insertion): For any pair of nodes $u, v \in V$ s.t. $(u, v) \notin E^*$, we have $(u, v) \in E$ with probability q

Illustration



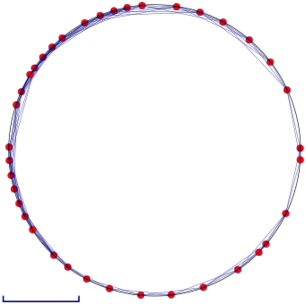
Hidden domain M

Illustration



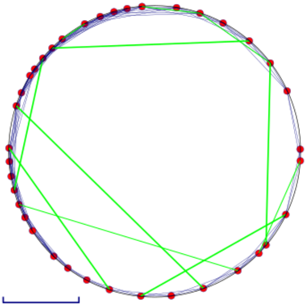
Graph Nodes V

Illustration



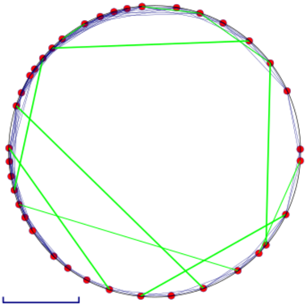
True graph G^*

Illustration



Random perturbation G

Illustration



Random perturbation G

The goal

Recover the shortest path metric d_{G^*} from G with approximation guarantee.

Remarks

- In many graphs, e.g social networks, nodes sampled from a hidden feature space, and edges encode proximity between graph nodes in certain feature space.

Remarks

- In many graphs, e.g social networks, nodes sampled from a hidden feature space, and edges encode proximity between graph nodes in certain feature space.
- Sampling from a measure allows varing degree distribution

Remarks

- In many graphs, e.g social networks, nodes sampled from a hidden feature space, and edges encode proximity between graph nodes in certain feature space.
- Sampling from a measure allows varing degree distribution
- Random Erdös-Rényi type perturbation allows exceptions / noise

Remarks

- In many graphs, e.g social networks, nodes sampled from a hidden feature space, and edges encode proximity between graph nodes in certain feature space.
- Sampling from a measure allows varing degree distribution
- Random Erdös-Rényi type perturbation allows exceptions / noise
- Shortest path metric natural choice in many situations (especially for sparse graphs), reflects the metric of the feature space
  - Other graph-induced metrics, e.g, diffusion distance?

Remarks

- In many graphs, e.g social networks, nodes sampled from a hidden feature space, and edges encode proximity between graph nodes in certain feature space.
- Sampling from a measure allows varing degree distribution
- Random Erdös-Rényi type perturbation allows exceptions / noise
- Shortest path metric natural choice in many situations (especially for sparse graphs), reflects the metric of the feature space
  - Other graph-induced metrics, e.g, diffusion distance?
- However, shortest path metric sensitive to random perturbations (especially "short-cuts")

Further remarks

- The model related to superposing a "structured subgraph" and a "random subgraph"
  - e.g, [Bollobás and Chung, 1988], [Watts and Strogatz, 1998], [Kleinberg 2000] (the small-world phenomenon), ...

Further remarks

- The model related to superposing a “structured subgraph” and a “random subgraph”
  - e.g, [Bollobás and Chung, 1988], [Watts and Strogatz, 1998], [Kleinberg 2000] (the small-world phenomenon), ...
- However, the metric recovery problem is somewhat orthogonal to goals in typical network analysis

Assumptions

Definition (Doubling measure)

A measure $\mu : X \to \mathbb{R}^+$ on a metric space (X, d) is said to be *L-doubling* if all metric balls have finite and positive measure and that there is a constant L such that for all $x \in X$ and $R > 0$,

$$\mu(B(x, 2R)) \leq L \cdot \mu(B(x, R)).$$

We call L the *doubling constant*.

Assumptions

Definition (Doubling measure)

A measure $\mu : X \to \mathbb{R}^+$ on a metric space (X, d) is said to be *L-doubling* if all metric balls have finite and positive measure and that there is a constant L such that for all $x \in X$ and $R > 0$,

$$\mu(B(x, 2R)) \leq L \cdot \mu(B(x, R)).$$

We call L the *doubling constant*.

Assumption-R: The parameter r (neighborhood size) is large enough such that $\mu(B(x, \frac{r}{2})) \geq s \geq \frac{12 \ln n}{n}$ for any $x \in M$.

Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Since $s \geq 12\ln n/n$ by Assumption-R, $p < \frac{1}{2e^{3/4}}$. As s increases, the upper bound on p gets closer to $1/2$.

Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.

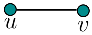
Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.



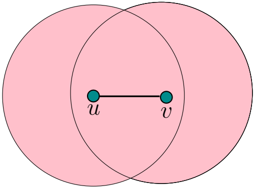
Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.



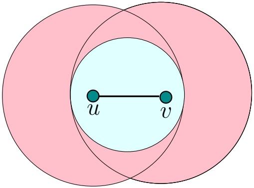
Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.



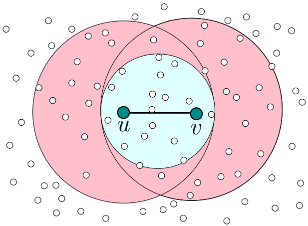
Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.



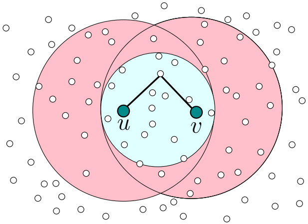
Effect of Deletion

Theorem (Deletion only)

Let G^* be the true graph generated as described, and G a graph obtained by deleting each edge in G^* with probability p. Assuming Assumption-R, *then for $p < \frac{1}{2}e^{-\frac{2\ln n}{sn}}$ with probability at least $1 - \frac{1}{n^{\Omega(1)}}$, the shortest path metric d_G in the observed graph is a 2-approximation of the shortest path metric d_{G^*} in the true graph; that is,*

$$\frac{1}{2}d_G(u,v) \leq d_{G^*}(u,v) \leq 2d_G(u,v).$$

Suppose an edge $(u, v) \in E^*$ is deleted in the observed graph G.

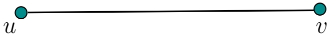


Effect of Insertion

Consider a *very-bad* inserted edge $(u, v) \in E$, meaning that $d_{G^*}(u, v) > 2$.

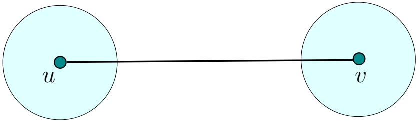
Effect of Insertion

Consider a *very-bad* inserted edge $(u, v) \in E$, meaning that $d_{G^*}(u, v) > 2$.



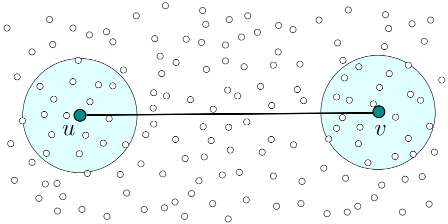
Effect of Insertion

Consider a *very-bad* inserted edge $(u, v) \in E$, meaning that $d_{G^*}(u, v) > 2$.



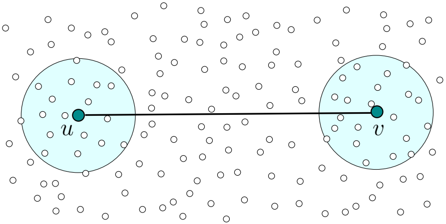
Effect of Insertion

Consider a *very-bad* inserted edge $(u, v) \in E$, meaning that $d_{G^*}(u, v) > 2$.



Effect of Insertion

Consider a *very-bad* inserted edge $(u, v) \in E$, meaning that $d_{G^*}(u, v) > 2$.



τ-Jaccard-Cleanup: Given graph G, for each edge $(u, v) \in G$, we keep the edge in a filtered graph $\widehat{G}$ iff

$$\rho_{u,v}(G) = \frac{|N_u^G \cap N_v^G|}{|N_u^G \cup N_v^G|} \geq \tau.$$

Insertion only – Good edges

Good edges have "large" Jaccard index.

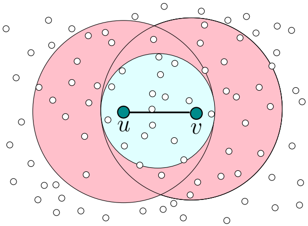
Lemma

Let V be n points sampled i.i.d. from L-doubling measure $\mu : M \to \mathbb{R}$. Let G^* be the r-neighborhood graph for V and $\widehat{G}$ obtained by inserting each edge not in G^* independently with probability q. Suppose Assumption-R *holds and the insertion probabiliy satisfies $q \leq cs$. Then w.h.p., for any $\tau \leq \frac{1}{(6+12c)L^2}$, $\rho_{u,v}(\widehat{G}) \geq \tau$ for all pairs of nodes $u, v \in V$ with $(u, v) \in E(G^*)$.*

- For example, if $c = \frac{1}{2}$ (i.e, $q \leq \frac{s}{2}$), then $\rho_{u,v}(\widehat{G}) \geq \frac{1}{13L^2}$ w.h.p.
- c can be super-constant, and tradeoff the requirement on q and Jaccard index on good edges.
    - As c increases, q is larger, but the upper bound on τ decreases.

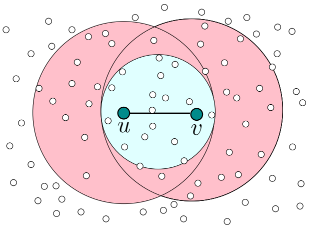
Requirement of $q \leq cs$

Recall the Jaccard index for an edge (u, v) is $\rho_{u,v}(G) = \frac{|N_u^G \cap N_v^G|}{|N_u^G \cup N_v^G|}$.



Requirement of $q \leq cs$

Recall the Jaccard index for an edge (u, v) is $\rho_{u,v}(G) = \frac{|N_u^G \cap N_v^G|}{|N_u^G \cup N_v^G|}$.

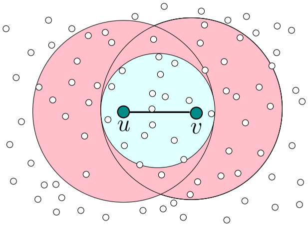


For an good edge $(u, v) \in E(G^*)$,

- When $q = 0$, $\rho_{u,v}(G)$ is a constant depending on L

Requirement of $q \leq cs$

Recall the Jaccard index for an edge (u, v) is $\rho_{u,v}(G) = \frac{|N_u^G \cap N_v^G|}{|N_u^G \cup N_v^G|}$.



For an good edge $(u, v) \in E(G^*)$,

- When $q = 0$, $\rho_{u,v}(G)$ is a constant depending on L
- As q increases, randomly inserted edges dominates, and $\rho_{u,v}(G)$ tends to q
  - $|N_u^G \cap N_v^G| \to nq^2$ while $|N_u^G \cup N_v^G| \to nq$

Insertion only – Bad edges

Very-bad edges have "small" Jaccard index.

Lemma

Let V be n points sampled i.i.d. from L-doubling measure μ. Let G^* be the r-neighborhood graph for V and $\widehat{G}$ obtained by inserting each edge not in G^* independently with probability q. Suppose Assumption-R *holds and the insertion probabiliy satisfies $q \leq cs$.*
Then for any $\tau \geq (c+2)q + 2(c+2)\sqrt{\frac{\ln n}{sn}}$, w.h.p., $\rho_{u,v}(\widehat{G}) < \tau$ for all pairs of nodes $u, v \in V$ such that (u, v) is very-bad.

- For example, if $c = 1$ an $sn = \omega(\ln n)$, then w.h.p. $\rho_{u,v}(\widehat{G}) \leq 3q + o(1)$ for all very-bad edges (u, v).

Main Result

Theorem

Given an observed graph G as a perturbed version of G^* as decribed before. Suppose Assumption-R holds, $sn = \omega(\ln n)$, the deletion probabily $p < \min\{1 - \frac{\sqrt{3}}{2}, \frac{1}{2}e^{-\frac{9 \ln n}{sn}}\}$, and that the insertion probability $q \leq cs$. Let $\widehat{G}_\tau$ denote the graph after τ-Jaccard-cleanup of G with $\tau \in (\frac{c}{1-p}q + o(1), \frac{2(1-p)^2}{15L^2(1+2c)})$. Then the shortest path distance metric $d_{\widehat{G}_\tau}$ from $\widehat{G}_\tau$ is a 2-approximation of the shortest path metric d_{G^*} of the true graph G^* with high probability.

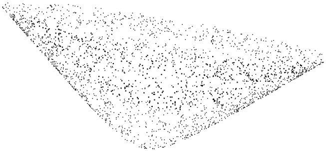
Main Result

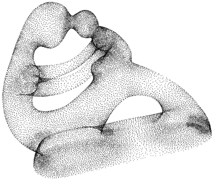
Theorem

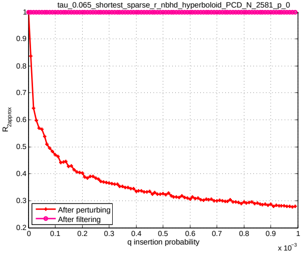
Given an observed graph G as a perturbed version of G^* as decribed before. Suppose Assumption-R holds, $sn = \omega(\ln n)$, the deletion probabily $p < \min\{1 - \frac{\sqrt{3}}{2}, \frac{1}{2}e^{-\frac{9 \ln n}{sn}}\}$, and that the insertion probability $q \leq cs$. Let $\widehat{G}_\tau$ denote the graph after τ-Jaccard-cleanup of G with $\tau \in (\frac{c}{1-p}q + o(1), \frac{2(1-p)^2}{15L^2(1+2c)})$. Then the shortest path distance metric $d_{\widehat{G}_\tau}$ from $\widehat{G}_\tau$ is a 2-approximation of the shortest path metric d_{G^*} of the true graph G^* with high probability.

- L-doubling measure can be extended to a local version

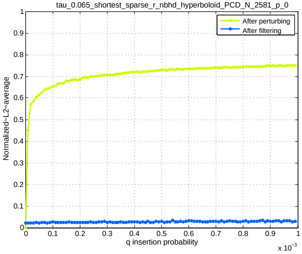
Preliminary Results – Proof of principle examples







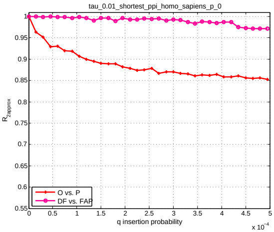
2-approximation rate



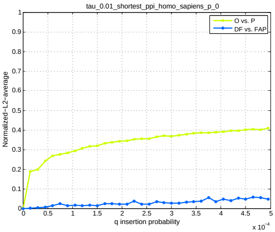
normalized L_2 error

Preliminary Results – Real networks w/o ground truth

- Given observed graph G, let G_q dentoe G with random $(p = 0, q)$-perturbation
- Let G^τ and G_q^τ be the graphs after τ-Jaccard filtering of G and G_q, respectively.
  - "O vs P": comparison between d_G and d_{G_q} as q increases
  - "DP vs FAP": comparison between d_{G^τ} and $d_{G_q^\tau}$



2-approximation rate



normalized L_2 error

Discussions

In this talk:

- Setting 1: towards parameter-free denoising for embedded point cloud data (PCD)
- Setting 2: metric embedding with outliers
- Setting 3: shortest path metric recovery from perturbed graphs

Discussions

In this talk:

- Setting 1: towards parameter-free denoising for embedded point cloud data (PCD)
- Setting 2: metric embedding with outliers
- Setting 3: shortest path metric recovery from perturbed graphs
- Other natural noise models?
  - E.g., for graph metrics, better tolerance in insertion probability, or better model to include more general graphs
  - for weighted graphs?

Discussions

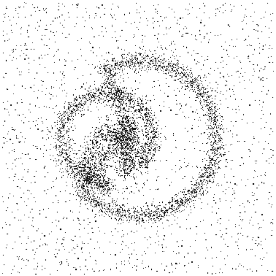
In this talk:

- Setting 1: towards parameter-free denoising for embedded point cloud data (PCD)
- Setting 2: metric embedding with outliers
- Setting 3: shortest path metric recovery from perturbed graphs

- Other natural noise models?
  - E.g., for graph metrics, better tolerance in insertion probability, or better model to include more general graphs
  - for weighted graphs?
- What are other ways to handle noise in metric?
  - Do we have to perform explicit denoising?

Bootstrapping?

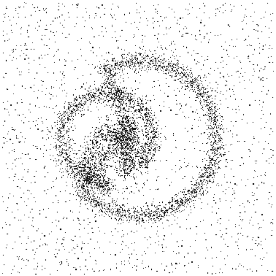
Given a "noisy" sample P of a hidden space already embedded



Bootstrapping?

Given a “noisy” sample P of a hidden space already embedded

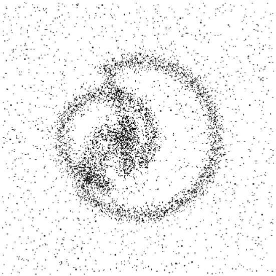
- Take multiple subsamples of P



Bootstrapping?

Given a "noisy" sample P of a hidden space already embedded

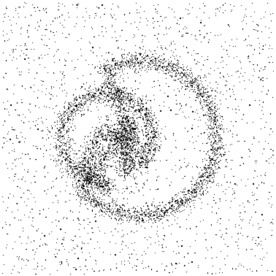
- Take multiple subsamples of P
- Compute persistence diagram for appropriated distance function (e.g, combined with distance to measure?)



Bootstrapping?

Given a "noisy" sample P of a hidden space already embedded

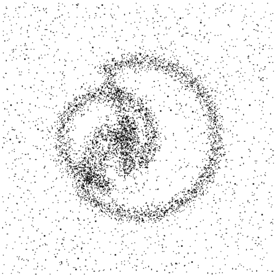
- Take multiple subsamples of P
- Compute persistence diagram for appropriated distance function (e.g, combined with distance to measure?)
- Average the set of resulting persistence diagrams



Bootstrapping?

Given a "noisy" sample P of a hidden space already embedded

- Take multiple subsamples of P
- Compute persistence diagram for appropriated distance function (e.g, combined with distance to measure?)
- Average the set of resulting persistence diagrams



Goal: depending on input noise model, develop theoretical guarantee for the output.

Bootstrapping?

This kind of bootstrapping idea appears more challenging for the (sparse) graph models like the one we introduced earlier.

Bootstrapping?

This kind of bootstrapping idea appears more challenging for the (sparse) graph models like the one we introduced earlier.

The geometry of the underlying space where graph nodes are sampled from may help.

Multiple Sets of Samples

What can we obtain if we are given multiple sets of samples of input data

- e.g, point sets $P_1, P_2, \ldots, P_k$ of a hidden domain

Multiple Sets of Samples

What can we obtain if we are given multiple sets of samples of input data

- e.g, point sets $P_1, P_2, \ldots, P_k$ of a hidden domain
- graph case?

Multiple Sets of Samples

What can we obtain if we are given multiple sets of samples of input data

- e.g, point sets $P_1, P_2, \ldots, P_k$ of a hidden domain
- graph case?
  - Averaging resulting persistence diagrams may not "cancel" noise.
  - Maybe "decorated" persistence diagrams?