A Quest to Unravel the Metric Structure Behind Perturbed Networks*†

Srinivasan Parthasarathy[1], David Sivakoff[2], Minghao Tian[3], and Yusu Wang[4]

1 **Computer Science and Engineering Department, The Ohio State University, Columbus, OH, USA**
srini@cse.ohio-state.edu

2 **Statistics and Mathematics Departments, The Ohio State University, Columbus, OH, USA**
dsivakoff@stat.osu.edu

3 **Computer Science and Engineering Department, The Ohio State University, Columbus, OH, USA**
tian.394@osu.edu

4 **Computer Science and Engineering Department, The Ohio State University, Columbus, OH, USA**
yusu@cse.ohio-state.edu

Abstract

Graphs and network data are ubiquitous across a wide spectrum of scientific and application domains. Often in practice, an input graph can be considered as an observed snapshot of a (potentially continuous) hidden domain or process. Subsequent analysis, processing, and inferences are then performed on this observed graph. In this paper we advocate the perspective that an observed graph is often a noisy version of some discretized 1-skeleton of a hidden domain, and specifically we will consider the following natural network model: We assume that there is a true graph G^* which is a certain proximity graph for points sampled from a hidden domain $\mathcal{X}$; while the observed graph G is an Erdös-Rényi type perturbed version of G^*.

Our network model is related to, and slightly generalizes, the much-celebrated small-world network model originally proposed by Watts and Strogatz. However, the main question we aim to answer is orthogonal to the usual studies of network models (which often focuses on characterizing / predicting behaviors and properties of real-world networks). Specifically, we aim to recover the metric structure of G^* (which reflects that of the hidden space $\mathcal{X}$ as we will show) from the observed graph G. Our main result is that a simple filtering process based on the *Jaccard index* can recover this metric within a multiplicative factor of 2 under our network model. Our work makes one step towards the general question of inferring structure of a hidden space from its observed noisy graph representation. In addition, our results also provide a theoretical understanding for Jaccard-Index-based denoising approaches.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, G.2.2 Graph Theory

Keywords and phrases metric structure, Erdös-Rényi perturbation, graphs, doubling measure

Digital Object Identifier 10.4230/LIPIcs.SoCG.2017.53

[*] A full version of the paper is available at https://arxiv.org/abs/1703.05475.

[†] This work is in part supported by National Science Foundation under grants IIS-1550757 and CCF-1618247. SP and DS would like to acknowledge NSF grant #DMS: 1418265 for partially supporting this work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

33rd International Symposium on Computational Geometry (SoCG 2017).
Editors: Boris Aronov and Matthew J. Katz; Article No. 53; pp. 53:1–53:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Graphs and networks are ubiquitous across a wide spectrum of scientific and application domains. Analyzing various types of graphs and network data play a fundamental role in modern data science. In the past several decades, there has been a large amount of research studying various aspects of graphs, ranging from developing efficient algorithms to process graphs, to information retrieval and inference based on graph data.

In many cases, we can view an input graph as an observed (discrete) 1-skeleton of a (potentially continuous) hidden domain. Subsequent analysis, processing, and inferences are then performed on this observed graph, with the ultimate goal being to understand the hidden space where the graph is sampled from. Many beautiful generative models for graphs have been proposed [9, 20], aiming to understand this transition process from a hidden space to the observed 1-skeleton, and to facilitate further tasks performed on graphs.

One line of such generative graph models assumes that an observed network is obtained by adding random perturbation to a specific type of underlying "structured graph" (such as a grid or a ring). For example, the much-celebrated small-world model by Watts and Strogatz [26] generates a graph by starting with a k-nearest neighbor graph spanned by nodes regularly distributed along a ring. It then randomly "rewires" some of the edges connecting neighboring points to instead connect nodes possibly far away. Watts and Strogatz showed that this simple model can generate networks that possess features of both a random graph and a proximity graph, and display two important characteristics often seen in real networks: low diameter in shortest path metric and high clustering coefficients. There have since been many variants of this model proposed so as to generate networks with different properties, such as adding random edges in a distance-dependent manner [23, 15], or extending similar ideas to incorporate hierarchical structures in networks; e.g, [16, 25]. There have also been numerous studies on characterizing statistical summaries, such as the average path lengths or the degree distributions, of small-world like networks; e.g [5, 11]; see [24, 6] for surveys.

Our work. In this paper, we take the perspective that an observed graph can be viewed as a noisy snapshot of the discretized 1-skeleton of a hidden domain of interest, and propose the following network model: Assume that the hidden space that generates data is a "nice" measure μ supported on a compact metric space $\mathcal{X} = (X, d_X)$ (e.g, the uniform measure supported on an embedded smooth low-dimensional Riemannian manifold). Suppose that the data points V are sampled i.i.d from this measure μ, and the "true graph" G_r^* connecting them is the r-neighborhood graph spanned by V (i.e, two points u, v are connected if their distance $d_X(u, v) \leq r$). The observed graph G however is only a noisy version of the true proximity graph G_r^*, and we model this noise by an Erdös-Rényi (ER) type perturbation – each edge in the true graph G_r^* can be deleted with probability p, while a "short-cut" edge between two unconnected nodes u, v could be inserted to G with probability q.

To motivate this model, imagine in a social network a person typically makes friends with other persons that are close to herself in the unknown feature space modeled by our metric space $\mathcal{X}$. The distribution of people (graph nodes) is captured by the measure μ on $\mathcal{X}$. However, there are always (or may be even many) exceptions – friends could be established by chance, and two seemingly similar persons (say, close geographically and in tastes) may not develop friendship. Thus it is reasonable to model an observed social network G as an ER-type perturbation of the proximity graph G_r^* to account for such exceptions.

The general question we hope to address is how to recover various properties of the hidden domain $\mathcal{X}$ from the observed graph G. In this paper we investigate a specific problem: how to recover the metric structure of G^*_r (induced by the shortest path distances in G^*_r) from the noisy observation G. As we show in Theorem 5, the metric structure of G^*_r "approximates" that of the hidden domain $\mathcal{X}$. Note that a few inserted "short-cuts" could significantly change the shortest path metric, one potential factor leading to the small-world phenomenon. Our main result is that a simple filtering procedure based on the so-called *Jaccard index* can recover the shortest path metric of G^*_r within a multiplicative factor of 2 (with high probabilities). We also provide some preliminary experimental results.

Remarks and discussion. The problem of recovering G^*_r from the observed graph G is different and orthogonal to the usual studies on similar network models: Those studies often focus on characterizing the graphs generated by such models and whether those characteristics match with real networks. We instead aim to recover metric structure of a hidden true graph G^*_r from a given graph G. There are different motivations for this task. For example, it could be that the true graph G^*_r is the real object of interest, and we wish to "denoise" the observed graph G to get a more accurate representation of G^*_r. Indeed, in [12], Godberg and Roth empirically show how to use small-world model to help remove false edges in protein-protein interaction (PPI) networks. See [4] for more examples.

Furthermore, even if the observed graph G is of interest itself, we may still want to recover information about the domain $\mathcal{X}$ where G is generated from. For example, suppose we are given two networks G_1 and G_2 modeling say the collaboration networks from two different disciplines, and our goal is to compare the hidden collaboration structures behind the two disciplines. Comparing the precise graph structures of observed graphs G_1 and G_2 could be misleading, as even if they are generated from the same hidden space $\mathcal{X}$, they could still look different due to the random generation process. It is more robust if we can compare the two hidden spaces generating them instead.

Finally, we remark that similar to the small-world network models, our model also overlays a random perturbation over a "structured" network. Indeed, our network model in some sense generalizes the small-world network model by Watts and Strogaz. Specifically, in the model by Watts and Strogaz (and some later variants), the underlying "structured" network is a ring (or lattice). In our case, we assume that graph nodes P are sampled from a measure μ and using the r-neighborhood proximity graph G^*_r to model this underlying "structured" network. This setup adds generality to our model: For example, it allows us to produce non-uniform and more complex degree distributions than those previously produced by starting with lattice vertices. At the same time, by putting conditions on the measure μ, it still gives us sufficient structure to relate G^*_r and G, as we will show in this paper. We also point out that the theoretical results hold for graphs across a range of density, where the number of edges could range from $\Theta(n \log n)$ to $\Theta(n^2)$.

All missing proofs due to lack space can be found in the full version [19].

2 Model for Perturbed Network

We now introduce a general model to generate an observed network G. Suppose we are given a compact geodesic metric space $\mathcal{X} = (X, d_X)$ [1][7]. Intuitively, we view an observed graph $G = (V, E)$ as a noisy 1-skeleton of $\mathcal{X}$, where graph nodes V of G are sampled from this

[1] A geodesic metric space is a metric space where any two points in it are connected by a path whose length equals the distance between them. Riemannian manifolds or compact sets in the Euclidean space are all geodesic metric spaces.

hidden metric space. More precisely, we will assume that V is sampled i.i.d. from a measure $\mu : X \to \mathbb{R}^+$ supported on X.

▶ **Definition 1** (Measure). Given a topological space X, a *measure* μ on X is simply a function that maps every Borel subset B of X to a non-negative number $\mu(B)$ which is additive: that is the measure of a countable family of pairwise-disjoint Borel subsets of X equals the sum of their respective measures.

In this paper, a measure is always a *probability measure*, meaning that $\mu(X) = 1$. To provide sufficient structure to the observed graph G so that it is not completely arbitrary, we want to assert some reasonable conditions on μ. To this end, we consider doubling measures:

▶ **Definition 2** (Doubling measure [13]). Given a metric space $\mathcal{X} = (X, d_X)$, let $\mathrm{B}(x, r) \subset X$ denotes the open metric ball $\mathrm{B}(x, r) = \{y \in X \mid d_X(x, y) < r\}$. A measure μ on $\mathcal{X}$ is said to be *doubling* if balls have finite and positive measure and there is a constant $L = L(\mu)$ s.t. for all $x \in X$ and any $r > 0$, we have $\mu(\mathrm{B}(x, 2r)) \leq L \cdot \mu(\mathrm{B}(x, r))$. We call L the *doubling constant* and say μ is an *L-doubling measure.*

These conditions on the measure also implies conditions on the underlying space X supporting the measure. Specifically, it is known that any metric space supporting a doubling measure has to be doubling as well, with its doubling constant depending on that of the measure [13].

Network model. We now describe our network model. Given a compact metric space $\mathcal{X} = (X, d_X)$ and an L-doubling measure $\mu : X \to \mathbb{R}^+$ supported on X, let V be a set of n points sampled i.i.d. from μ. We assume that the *true graph* $G^*_r = (V, E^*)$ is the r-neighborhood graph for some parameter $r > 0$; that is, $E(G^*) = E^* = \{(u, v) \mid d_X(u, v) \leq r, u, v \in V\}$.

▶ **Definition 3.** The *observed graph* $G(r, p, q) = (V, E)$ is based on $G^*_r = (V, E^*)$, but with the following two types of random perturbations:

- p-deletion: For each edge $(u, v) \in E^*$, (u, v) is in the observed graph $G(r, p, q)$ with probability $1 - p$ (that is, an edge in E^* is deleted with probability p).
- q-insertion: For any pair of nodes $u, v \in V$ s.t. $(u, v) \notin E^*$, we have that $(u, v) \in E$ with probability q.

Intuitively, in our model, the observed network G is a random geometric graph sampled from the metric space $\mathcal{X}$ which then undergoes Erdös-Rényi type perturbation. In what follows, we often omit the parameters r, p, q from the notations G^*_r and $G(r, p, q)$, when their choices are clear from the context. Note that both G^* and G are unweighted graphs (that is, all edges have weight 1). We now equip each graph with its shortest path metric, and obtain two discrete metric space (V, d_{G^*}) and (V, d_G) induced by G^* and G, respectively.

Problem statement and main results. Adding short-cuts (via q-insertions) could significantly distort the shortest path metric in G^*. Our ultimate goal is to infer information about both $\mathcal{X}$ and μ where points are sampled from, through the study of the observed graph G. In this paper we aim to recover the metric structure of G^* (as a reflection of metric structure of $\mathcal{X}$) from G. Specifically, we show that a simple filtering process based on the so-called Jaccard index can remove sufficient "bad edges" in G so as to recover the shortest path metric of G^* up to a factor of 2 w.h.p.

▶ **Definition 4** (Jaccard index). Given an arbitrary graph G, let $N_G(u)$ denote the set of neighbors of u in G (i.e. nodes connected to $u \in V(G)$ by edges in $E(G)$). Given any edge

$(u, v) \in E(G)$, the *Jaccard index* $\rho_{u,v}$ of this edge is defined as

$$\rho_{u,v}(G) = \frac{|N_G(u) \cap N_G(v)|}{|N_G(u) \cup N_G(v)|}. \tag{1}$$

We remark that Jaccard index is a popular way to measure similarity between a pair of nodes connected by an edge in a graph [17], and has been commonly used in practice for denoising and sparsification purposes [22, 21]. Our results provide a theoretical understanding for such empirical Jaccard-based denoising approaches.

The main result is stated in Theorem 13. To show how this is established, we show two results on the influence of the shortest path under the p-deletion (Theorem 9) and under the q-insertion (Theorem 12), respectively. The proof for Theorem 13 combines the ideas for proofs of these two results.

Metric structures for G^*_r versus for $\mathcal{X}$. Our main results recover the shortest path metric for G^*_r approximately. In some sense, the metric of a proximity graph provides an approximation of that of X, the domain where input graph nodes are sampled from; see e.g, [1, 8] for the case where X is a smooth Riemannian manifold embedded in Euclidean space.

We make this relationship precise for our setting as follows. The proof of this result is rather standard (see e.g, the proof of Theorem 5.2 of [8]) and can be found in the full version [19].

▶ **Theorem 5.** *Let (X, d_X) be a compact geodesic metric space and μ a doubling measure supported on X. Let V_n be a set of n points sampled i.i.d. from μ, and G^*_r the r-neighborhood graph constructed on V_n (each edge in G^*_r has equal weight 1) with the associated shortest path metric $d_{G^*_r}$. For any sample V_n, consider the distance between $r \cdot d_{G^*_r}$ ($d_{G^*_r}$ scaled by r) and d_X restricted to the sample V_n; that is,*

$$\|r \cdot d_{G^*_r} - d_X|_{V_n}\|_\infty := \max_{v, v' \in V_n} |r \cdot d_{G^*_r}(v, v') - d_X(v, v')|.$$

Then we have that for a fixed r, $\limsup_{n\to\infty} \|r \cdot d_{G^*_r} - d_X|_{V_n}\|_\infty \le r$ almost surely.

3 Recovering the shortest path metric of G^*

To illustrate the main idea, we first consider the deletion-only and insertion-only perturbation of the true graph G^* in Sections 3.1 and 3.2, respectively. As we will see below, the main difficulty lies in handling insertions (short-cuts). We then combine the two cases and present our main result, Theorem 13. First, we describe one (natural) assumption on r that we will use later in all our statements.

Note that as r tends to 0, the corresponding r-neighborhood graph may be very sparse, and a sparse graph G^*_r is quite sensitive to random deletions and insertions. We would like to consider r in a range where is meaningful. We make the following assumption, asserting a lower-bound on the mass contained inside any metric ball of radius $r/2$:

[Assumption-R]: The parameter r is large enough such that for any $x \in X$, $\mu(\mathrm{B}(x, \frac{r}{2})) \ge \mathrm{s}$ where s satisfies $\mathrm{s} \ge \frac{12 \ln n}{n-2} (= \Omega(\frac{\ln n}{n}))$.

Intuitively, r is large enough such that with high probability each vertex v in G^*_r has degree $\Omega(\ln n)$. Note that requiring r to be large enough to have an $\Omega(\ln n/n)$ lower bound on the measure of any metric ball is natural. For example, for a random geometric graph $G(r, n)$ constructed as the r-neighborhood graph for points i.i.d. sampled from a uniform measure on a Euclidean cube, asymptotically this is the same requirement so as to make sure that the resulting r-neighborhood graph is connected with high probability [20].

The proof of the following observation is simple and can be found in [19].

▶ **Lemma 6.** *Under Assumption-R, with probability at least* $1-n^{-5/3}$*, all vertices in* G_r^* *have more than* $\frac{s(n-1)}{3} > 4\ln n$ *neighbors.*

Since μ is a doubling measure, any two neighbors (u, v) in the r-neighborhood graph G_r^* would share many neighbors. Specifically, if (u, v) is an edge in G_r^*, that is, $d_X(u, v) \le r$, then $\mathrm{B}(u, r) \cap \mathrm{B}(v, r)$ must contain a metric ball of radius $r/2$ (say centered at midpoint z of a shortest path connecting u to v in X; see Figure 1 (a)). Thus by a similar argument as the proof of Lemma 6, we obtain the following bound on the number of common neighbors between the nodes u, v if edge $(u, v) \in G_r^*$.

▶ **Corollary 7.** *Assume that the graph nodes* V *of* G_r^* *are sampled i.i.d from an* L*-doubling measure* μ *supported on a compact geodesic metric space* (X, d_X)*. Then under Assumption-R, with probability at least* $1-n^{-2/3}$*, any two neighbors* $(u, v) \in G_r^*$ *have* $\frac{s(n-1)}{3} > 4\ln n = \Omega(\ln n)$ *number of common neighbors.*

3.1 Deletion only

In this case, we assume that we remove each edge in G^* independently with probability p to obtain an observed empirical graph $\widehat{G}$. Our goal is to relate the shortest path metrics d_{G^*} of G^* and $d_{\widehat{G}}$ of $\widehat{G}$ respectively. Deletion-only means that shortest path distances in $\widehat{G}$ are larger than those in G^*. Since any two nodes u, v connected in G^* share sufficient number ($\Omega(\ln n)$) of common neighbors, intuitively, removing even a constant fraction of edges in G^* can still guarantee that w.h.p. u and v will still have some common neighbors left, and thus u and v can be connected through that common neighbor by a path of length 2 in $\widehat{G}$. Hence overall, w.h.p. the distortion in shortest path distance is at most by a factor of 2.

▶ **Definition 8.** Let G and G' be two graphs spanned on the same set of nodes V, and equipped with graph shortest path metric d_G and $d_{G'}$, respectively. By $d_G \le c d_{G'}$, we mean that for any two nodes $u, v \in V$, we have that $d_G(u, v) \le c d_{G'}(u, v)$. We say that $d_{G'}$ is *a* c*-approximation of* d_G if $\frac{1}{c} d_G \le d_{G'} \le c d_G$.

▶ **Theorem 9** (Random deletion)**.** *Let* V *be* n *points sampled i.i.d. from a probability measure* $\mu : X \to \mathbb{R}^+$ *supported on a compact metric space* (X, d_X)*. Let* G^* *be the* r*-neighborhood graph for* V*; and* $\widehat{G}$ *a graph obtained by removing each edge in* G^* *independently with probability* p*. Under Assumption-R and for* $p < \frac{1}{2} e^{-\frac{9\ln n}{s(n-1)}}$*, we have with probability at least* $1 - \frac{1}{n^{\Omega(1)}}$*, the shortest path metric* $d_{\widehat{G}}$ *is a 2-approximation of the shortest path metric* d_{G^*}*.*

Specifically, since $s > \frac{12\ln n}{n-1}$*, the statement holds for* $p < \frac{1}{2e^{3/4}}$*. As* s *becomes larger, the upper bound on* p *gets closer to* $1/2$*.*

Proof. For a node $u \in V$, let $N_{G^*}(u)$ and $N_{\widehat{G}}(u)$ denote the set of neighbors of u in graph G^* and graph $\widehat{G}$, respectively.

Since deletion cannot decrease the length of shortest paths, we have $d_{G^*} \le d_{\widehat{G}}$. We now show that $d_{\widehat{G}} \le 2 d_{G^*}$.

Consider $(u, v) \in E(G^*)$: Assume that they share $k_{u,v}$ number of common neighbors; that is, $k_{u,v} = |N_{G^*}(u) \cap N_{G^*}(v)|$. The probability that $N_{\widehat{G}}(u) \cap N_{\widehat{G}}(v) = \emptyset$ (i.e, u and v have no common neighbor in graph $\widehat{G}$) is thus $(2p)^{k_{u,v}}$.

On the other hand, by Corollary 7, with probability at least $1 - n^{-2/3}$ we have that $k_{u,v} \geq s(n-1)/3$ for all $(u, v) \in E(G^*)$. By applying the law of total probability, it then follows that the probability that there exists any $(u, v) \in E(G^*)$ with $N_{\widehat{G}}(u) \cap N_{\widehat{G}}(v) = \emptyset$ is at most: $n^{-2/3} + n^2(2p)^{s(n-1)/3} < n^{-2/3} + n^2(e^{-3\ln n}) < n^{-1/3}$, where we plug in the bound on p to derive the first inequality.

Hence with probability at least $1 - n^{-1/3}$, we have that for all edges $(u, v) \in E(G^*)$, their distance in $\widehat{G}$ satisfies $d_{\widehat{G}}(u, v) \leq 2$ (via one of their common neighbor in $N_{\widehat{G}}(u) \cap N_{\widehat{G}}(v)$). This in turn implies that with probability at least $1 - n^{-1/3}$, for any path $\pi = \langle v_1, \dots, v_m\rangle$ in G^* with length m, we can find a path of length at most $2m$ in $\widehat{G}$ to connect v_1 to v_m (as each edge (v_i, v_{i+1}) in π corresponds to a path of length at most 2 in $\widehat{G}$). If u and v are disconnected in G^*, then obviously they are still disconnected in $\widehat{G}$. Hence, for any two $u, v \in V$, $d_{\widehat{G}}(u, v) \leq 2d_{G^*}(u, v)$, and the theorem follows. ◀

3.2 Insertion only

Now assume that the observed graph $\widehat{G}$ is generated from the true graph G^* where all edges in G^* also exist in $\widehat{G}$, and for any $u, v \in V$ with $(u, v) \neq E(G^*)$, we have $(u, v) \in E(\widehat{G})$ with probability q. In this case, the shortest path metric can be significantly altered in $d_{\widehat{G}}$. Hence to recover the metric d_{G^*}, instead of operating on $\widehat{G}$ directly, we will construct another graph $\tilde{G}$ from $\widehat{G}$, so that its shortest path metric $d_{\tilde{G}}$ approximates d_{G^*}.

We propose the following Jaccard-Index-based filtering process, which we call a τ-Jaccard filtering, as it uses a parameter τ. (Recall the definition of Jaccard index in Def. 4). We represent the output filtered (denoised) graph as $\tilde{G}_\tau$:

τ-Jaccard filtering: Given graph $\widehat{G}$, for each edge $(u, v) \in E(\widehat{G})$, we insert the edge (u, v) into $E(\tilde{G}_\tau)$ if and only if $\rho_{u,v}(\widehat{G}) \geq \tau$. That is, $V(\tilde{G}_\tau) = V(\widehat{G})$ and $E(\tilde{G}_\tau) := \{(u, v) \in E(\widehat{G}) \mid \rho_{u,v}(\widehat{G}) \geq \tau\}$.

Below we first show that w.h.p., all "good" edges in the true r-neighborhood graph G^* will have a large Jaccard index, so that they will be kept in $\tilde{G}_\tau$ after a τ-Jaccard filtering procedure with appropriate τ.

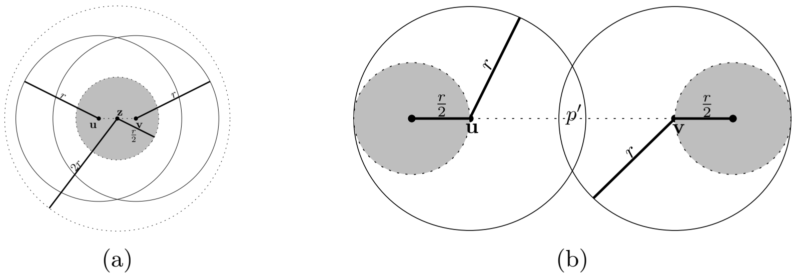
▶ **Lemma 10.** *Let V be a set of n points sampled i.i.d. from an L-doubling probability measure μ supported on a compact geodesic metric space $\mathcal{X} = (X, d_X)$. If Assumption-R holds and $q \leq cs$, then for $\forall\tau \leq \frac{1}{(6+\frac{1}{\ln n}+12c)L^2}$, we have with probability at least $1 - n^{-2/3}$, that $\rho_{u,v}(\widehat{G}) \geq \tau$ for all pairs of nodes $u, v \in V$ with $d_X(u, v) \leq r$.*

For example, if $c = \frac{1}{2}$ (i.e, $q \leq \frac{s}{2}$), then the bound on $\rho_{u,v}$ holds for $\tau \leq \frac{1}{13L^2}$. Note c may not be a constant and can depend on n; as c increases, the upper bound on τ decreases.

Proof. Consider a fixed pair of nodes $u, v \in V$, and let $F = F(u, v)$ be the event that $d_X(u, v) \leq r$. Set $\alpha_* = |N_{G^*}(u) \cap N_{G^*}(v)|$ to be the number of common neighbors of u and v in G^*. Let $\beta = |N_{\widehat{G}}(u) \cup N_{\widehat{G}}(v)|$ denote the total number of neighbors of u and v in the perturbed graph $\widehat{G}$.

Since $\widehat{G}$ can have only more edges than G^*, $|N_{\widehat{G}}(u) \cap N_{\widehat{G}}(v)| \geq |N_{G^*}(u) \cap N_{G^*}(v)| = \alpha_*$ and thus $\rho_{u,v}(\widehat{G}) \geq \frac{\alpha_*}{\beta}$. In what follows, we prove that $\frac{\alpha_*}{\beta} \geq \tau \cdot I_F$ (which implies that $\rho_{u,v}(\widehat{G}) \geq \tau \cdot I_F$) with probability at least $1 - 2n^{-8/3}$. (Here, we use I_A to denote the indicator random variable of the event A, and the conventions that $\rho_{u,v}(\widehat{G}) = 0$ if $(u, v) \notin \widehat{G}$ and $0/0 = 0$.)

Note that α_* is a random variable, which equals the number of (i.i.d. sampled) points from $V - \{u, v\}$ that fall in the region $\mathrm{B}(u, r) \cap \mathrm{B}(v, r)$. That is, conditional on u and v, α_*



(a) (b)

Figure 1 In these figures, we draw metric balls as Euclidean balls just for illustration purpose. (a) illustrates the bound $p_{\alpha*} \geq \mu(\mathrm{B}(z, r/2))$ which follows from $\mathrm{B}(z, r/2) \subseteq \mathrm{B}(u, r) \cap \mathrm{B}(v, r)$. (b) Key observation for Lemma 11: as $d_X(u, v) > r$, we have that the region $[B(u, r) \cup B(v, r)] \setminus [B(u, r) \cap B(v, r)]$ contains at least two metric balls, each of radius $r/2$.

is drawn from a binomial distribution $Bin(n-2, p_{\alpha_*})$ with $p_{\alpha_*} = \mu(\mathrm{B}(u, r) \cap \mathrm{B}(v, r))$, and the conditional expectation of α_* given u and v is $\delta_{\alpha_*} = (n-2) \cdot p_{\alpha_*}$.

Now observe that, conditional on u and v, the random variable $\beta - 2$ (see footnote[2]) has distribution $Bin(n-2, p_\beta)$ with $p_\beta = p_{\beta_*} + (1-p_{\beta_*})(2q-q^2)$, where $p_{\beta_*} = \mu(\mathrm{B}(u, r) \cup \mathrm{B}(v, r))$. Indeed, observe that, conditional on u and v, points contributing to β can be generated as follows. Let $U = \mathrm{B}(u, r) \cup \mathrm{B}(v, r)$. Independently, for each $i = 1, \ldots, n-2$, we draw a point x_i randomly from μ and we also perform an independent coin flip for this point, with probability of heads equal to $1 - (1-q)^2 = 2q - q^2$. This quantity is the probability for a point *outside* U to be connected to either u or v under edge-insertion probability q. We set the indicator variable $y_i = 1$ iff either $x_i \in U$, or $x_i \notin U$ and the i^{th} coin flip is heads. Conditional on u and v, the resulting $n-2$ indicator random variables $y_1, \ldots, y_{n-2}$ are i.i.d. with $\mathrm{P}[y_i = 1 \mid u, v] = p_{\beta_*} + (1-p_{\beta_*})(2q-q^2) = p_\beta$. Therefore, given u and v, the distribution of $\beta - 2 = \sum y_i$ is $Bin(n-2, p_\beta)$. The conditional expectation of β given u and v, denoted δ_β, satisfies

$$(n-2) \cdot p_{\beta_*} \leq \delta_\beta = (n-2) \cdot p_\beta + 2 \leq (n-2) \cdot p_{\beta_*} + (n-2) \cdot 2q + 2. \tag{2}$$

Let us for now assume that $\frac{c_1 \delta_{\alpha_*}}{c_2 \delta_\beta} \geq \tau I_F$ a.s. for constants $c_1 = 1 - \sigma_1$ and $c_2 = 1 + \sigma_2$ with $0 < \sigma_1 < 1$ and $0 < \sigma_2$ to be set shortly.

If $d_X(u, v) \leq r$, then $\mathrm{B}(u, r) \cap \mathrm{B}(v, r)$ contains at least one metric ball of radius $r/2$ (say $\mathrm{B}(z, r/2)$ with z being the mid-point of a shortest path between u and v in $\mathcal{X}$; see Figure 1 (a)).

Hence by Assumption-R, on the event $d_X(u, v) \leq r$, we have

$$\delta_{\alpha_*} \geq (n-2) \cdot \mu(\mathrm{B}(z, r/2)) \geq (n-2) \cdot \frac{12 \ln n}{n-2} = 12 \ln n.$$

Similarly, using (2), the conditional expectation of β satisfies

$$\delta_\beta \geq (n-2) \cdot p_{\beta_*} \geq (n-2) \cdot \mu(\mathrm{B}(u, r)) \geq 12 \ln n. \tag{3}$$

We now set $\sigma_1 = 2/3$ and $\sigma_2 = 1$. It then follows from Chernoff bounds that

$$\mathrm{P}[\alpha_* < c_1 \delta_{\alpha_*} \mid u, v, F] + \mathrm{P}[\beta > c_2 \delta_\beta \mid u, v] \leq e^{-\frac{\sigma_1^2}{2} \delta_{\alpha*}} + e^{-\frac{\sigma_2}{3} \delta_\beta} \leq n^{-\frac{8}{3}} + n^{-4}.$$

[2] The subtraction of 2 in $\beta - 2$ accounts for points u and v, which are in $N_{\widehat{G}}(u) \cup N_{\widehat{G}}(v)$. Similarly, in the binomial distribution we will have only $n-2$, accounting for points in $V - \{u, v\}$.

Taking expectation of the above with respect to u and v gives

$$\mathrm{P}[\alpha_* < c_1\delta_{\alpha_*} \mid F] + \mathrm{P}[\beta > c_2\delta_\beta] \leq 2n^{-\frac{8}{3}}. \tag{4}$$

On the other hand, since $\frac{\alpha_*}{\beta} \geq 0$, we have

$$\begin{aligned}\mathrm{P}[\frac{\alpha_*}{\beta} < \tau I_F] \leq \mathrm{P}[\frac{\alpha_*}{\beta} < \tau \mid (\alpha_* \geq c_1\delta_{\alpha_*}) \wedge (\beta \leq c_2\delta_\beta) \wedge F] \\ + \mathrm{P}[(\{\alpha_* < c_1\delta_{\alpha_*}\} \vee \{\beta > c_2\delta_\beta\}) \wedge F].\end{aligned} \tag{5}$$

Since we assumed that $\frac{c_1\delta_{\alpha_*}}{c_2\delta_\beta} \geq \tau I_F$, if $\alpha_* \geq c_1\delta_{\alpha_*}$ and $\beta \leq c_2\delta_\beta$ and $d_X(u,v) \leq r$, then we have $\frac{\alpha_*}{\beta} \geq \frac{c_1\delta_{\alpha_*}}{c_2\delta_\beta} \geq \tau$. This means that

$$\mathrm{P}[\frac{\alpha_*}{\beta} < \tau \mid (\alpha_* \geq c_1\delta_{\alpha_*}) \wedge (\beta \leq c_2\delta_\beta) \wedge F] = 0.$$

Hence the first term in the right-hand side of (5) is 0. Together with (4), and recalling $\rho_{u,v}(\widehat{G}) \geq \frac{\alpha_*}{\beta}$, we have

$$\mathrm{P}[\rho_{u,v}(\widehat{G}) < \tau I_F] \leq \mathrm{P}[\frac{\alpha_*}{\beta} < \tau I_F] \leq 2n^{-\frac{8}{3}}.$$

By the union bound, the probability that $\rho_{u,v}(\widehat{G}) \geq \tau$ for all pairs of nodes $u, v \in V$ such that $d_X(u,v) \leq r$ is thus at least $1 - \frac{1}{2}n^2(2n^{-\frac{8}{3}}) = 1 - n^{-\frac{2}{3}}$.

Finally, we need to verify that $\frac{c_1\delta_{\alpha_*}}{c_2\delta_\beta} = \frac{\delta_{\alpha_*}}{6\delta_\beta} \geq \tau I_F$ holds for a.e. u and v. This holds automatically if $d_X(u,v) > r$, so assume $d_X(u,v) \leq r$. Recall that $\delta_\beta \leq (n-2)\cdot p_{\beta_*} + (n-2)\cdot 2q + 2$ by (2). Since $q \leq c\mathrm{s}$, we have $(n-2)2q \leq 2(n-2)c\mathrm{s}$. On the other hand, by Assumption-R, $p_{\beta_*} \geq \mu(\mathrm{B}(u,r)) \geq \mathrm{s}$, hence $2(n-2)q \leq 2(n-2)c \cdot p_{\beta_*}$. Combining this with the fact that $(n-2)p_{\beta_*} \geq 12\ln n$ from (3) (which also implies that $2 \leq \frac{(n-2)p_{\beta_*}}{6\ln n}$), it then follows that

$$\frac{\delta_{\alpha_*}}{6\delta_\beta} \geq \frac{\delta_{\alpha_*}}{6((n-2)(1+\frac{1}{6\ln n})p_{\beta_*} + 2(n-2)c\cdot p_{\beta_*})} = \frac{p_{\alpha_*}}{p_{\beta_*}} \cdot \frac{1}{6 + \frac{1}{\ln n} + 12c}. \tag{6}$$

Now let z be the midpoint of a geodesic connecting u and v; see Figure 1 (a). Observe that $p_{\alpha_*} \geq \mu(\mathrm{B}(z,r/2))$, $p_{\beta_*} \leq \mu(\mathrm{B}(z,2r))$ and since μ is L-doubling, we have:

$$p_{\beta_*} \leq \mu(\mathrm{B}(z,2r)) \leq L\mu(\mathrm{B}(z,r)) \leq L^2\mu(\mathrm{B}(z,r/2)) \leq L^2 p_{\alpha_*}. \tag{7}$$

Combining equations (6) and (7), we have that if $\tau \leq \frac{1}{(6+\frac{1}{\ln n}+12c)L^2}$, then $\frac{\delta_{\alpha_*}}{6\delta_\beta} \geq \tau$ is satisfied. This proves the lemma. ◀

Discussion on the bounds of parameters. Lemma 10 implies that, with high probability, we will not remove any good edges if the doubling constant L of the measure is at most $O(\frac{1}{\sqrt{\tau}})$ and the insertion probability is small ($q \leq c s$). The requirement that $L = O(\frac{1}{\sqrt{\tau}})$ is rather mild; we now inspect the requirement $q \leq cs$: Since sn lower-bounds the degree of a node in the true graph G^* (by Lemma 6), it is reasonable that the insertion probability q is required to be small compared to s; as otherwise, the "noise" (inserted edges) will overwhelm the signal (original edges). Furthermore, it is important to note that c is not necessarily a constant – it can depend on n, but as c increases, the upper bound of the admissible range for parameter τ decreases.

The following result complements Lemma 10 by stating that for insertion probability $q \leq cs$, all “really bad” edges in $\widehat{G}$ will have small Jaccard index, and thus will be removed by our τ-filtering process.

In particular, we define an edge $(u,v) \in E(\widehat{G}) \setminus E(G^*)$ in the observed graph $\widehat{G}$ to be *really-bad* if $N_{G^*}(u) \cap N_{G^*}(v) = \emptyset$. Note that $(u,v) \notin E(G^*)$ is equivalent to $d_X(u,v) > r$.

▶ **Lemma 11.** *Let V be a set of n points sampled i.i.d. from an L-doubling probability measure μ supported on a compact geodesic metric space $\mathcal{X} = (X, d_X)$. If Assumption-R holds and $q \leq \mathrm{cs}$, then $\forall \tau \geq (c+2)q + 2(c+2)\sqrt{\frac{\ln n}{s(n-2)}}$, we have with probability at least $1 - n^{-2}$, $\rho_{u,v}(\widehat{G}) < \tau$ for* all *pairs of nodes $u, v \in V$ such that (u,v) is really-bad.*

For example, if $c = 1$ and $\mathrm{s} \cdot n = \omega(\ln n)$, then the condition on τ is that $\tau \geq 3q + o(1)$.

Proof. Consider a fixed pair of nodes $u, v \in V$, and let $F = F(u,v)$ be the event that $N_{G^*}(u) \cap N_{G^*}(v) = \emptyset$ and $d_X(u,v) > r$. Let $\alpha = |N_{\widehat{G}}(u) \cap N_{\widehat{G}}(v)|$,

$$\alpha_I = \big|\{x \in N_{G^*}(u) \cup N_{G^*}(v) \,:\, x \text{ is connected to both } u \text{ and } v \text{ in } \widehat{G}\}\big|, \text{ and}$$
$$\alpha_o = \big|\{x \notin N_{G^*}(u) \cup N_{G^*}(v) \,:\, x \text{ is connected to both } u \text{ and } v \text{ in } \widehat{G}\}\big|.$$

Then we have $\alpha = \alpha_I + \alpha_o$. Set $\beta_* = |N_{G^*}(u) \cup N_{G^*}(v)|$, so we have $|N_{\widehat{G}}(u) \cup N_{\widehat{G}}(v)| \geq \beta_* + \alpha_o =: \beta$. It is easy to see that

$$\rho_{u,v}(\widehat{G}) = \frac{\alpha}{|N_{\widehat{G}}(u) \cup N_{\widehat{G}}(v)|} \leq \frac{\alpha}{\beta_* + \alpha_o} = \frac{\alpha}{\beta}.$$

We aim to show that with very high probability $\frac{\alpha}{\beta} I_F < \tau$, which implies that $\rho_{u,v}(\widehat{G}) I_F < \tau$.

First, we claim that, conditional on the locations of u and v and the event F, the distribution of α is $Bin(n-2, p_\alpha)$ with $p_\alpha = \frac{p_{\beta_*} - p'}{1-p'} q + \frac{1-p_{\beta_*}}{1-p'} q^2$, where $p_{\beta_*} = \mu(B(u,r) \cup B(v,r))$ and $p' = \mu(B(u,r) \cap B(v,r))$. We also claim that the conditional distribution of β given u, v and F is $Bin(n-2, p_\beta)$ with $p_\beta = \frac{p_{\beta_*} - p'}{1-p'} + \frac{1-p_{\beta_*}}{1-p'} q^2$. Details in [19].

If $d_X(u,v) > r$, the region $[B(u,r) \cup B(v,r)] \setminus [B(u,r) \cap B(v,r)]$ contains at least two disjoint metric balls of radius $r/2$; see Figure 1(b). Therefore, $p_{\beta_*} - p' \geq 2\mu(B(\frac{r}{2})) \geq 2s$. The conditional expectation of α given u, v and F, denoted by $\delta_\alpha (= (n-2)p_\alpha)$, satisfies:

$$(n-2)\frac{p_{\beta_*} - p'}{1-p'} q \leq \delta_\alpha = (n-2)[\frac{p_{\beta_*} - p'}{1-p'} q + \frac{1-p_{\beta_*}}{1-p'} q^2] \leq (1 + \frac{c}{2})(n-2)\frac{p_{\beta_*} - p'}{1-p'} q, \quad (8)$$

where the last inequality follows from $q \leq cs \leq c \cdot \frac{p_{\beta_*} - p'}{2}$. The conditional expectation of β given u, v and F, denoted δ_β, satisfies

$$\delta_\beta = (n-2)p_\beta \geq (n-2)\frac{p_{\beta_*} - p'}{1-p'}. \quad (9)$$

Let us now assume that $\frac{c_1 \delta_\alpha}{c_2 \delta_\beta} I_F \leq \tau$ a.s. for $c_1 = 1 + \epsilon$ and some constant $c_2 = 1 - \sigma$ with $\epsilon = \frac{2}{q}\sqrt{\frac{\ln n}{s(n-2)}}$ and some $0 < \sigma < 1$ to be set later.

If $q \leq 2\sqrt{\frac{\ln n}{s(n-2)}}$, then we have $\epsilon \geq 1$. In this case, combining Chernoff bounds with (8) and the fact that $p_{\beta_*} - p' \geq 2\mu(B(\frac{r}{2})) \geq 2s$ obtained earlier, we have

$$\begin{aligned} \mathrm{P}[\alpha \geq (1+\epsilon)\delta_\alpha \mid u, v, F] &\leq e^{-\frac{\epsilon}{3}\delta_\alpha} = e^{-\frac{2}{3q}\sqrt{\frac{\ln n}{s(n-2)}}\delta_\alpha} \leq e^{-\frac{2}{3q}\sqrt{\frac{\ln n}{s(n-2)}}(n-2)\frac{p_{\beta_*} - p'}{1-p'} q} \\ &\leq e^{-\frac{4}{3}\sqrt{(n-2)(\ln n)s}} \leq e^{-\frac{4}{3}\sqrt{(n-2)(\ln n)\frac{12 \ln n}{n-2}}} \leq n^{-4}. \end{aligned} \quad (10)$$

Otherwise, we have $q > 2\sqrt{\frac{\ln n}{s(n-2)}}$, so $0 < \epsilon < 1$. In this case, by Chernoff bounds

$$\begin{aligned} \mathrm{P}[\alpha \geq (1+\epsilon)\delta_\alpha \mid u,v,F] \leq e^{-\frac{1}{2}\epsilon^2\delta_\alpha} &\leq e^{-2\frac{\ln n}{\mathrm{s}(n-2)}\frac{1}{q^2}(n-2)\frac{p_{\beta_*}-p'}{1-p'}q} \\ &= e^{-2\frac{\ln n(p_{\beta_*}-p')}{\mathrm{s}q}} \leq e^{-2\ln n\cdot\frac{2\mathrm{s}}{\mathrm{s}q}} \leq n^{-4}. \end{aligned} \tag{11}$$

On the other hand, by Chernoff bounds, we have $\mathrm{P}[\beta \leq c_2\delta_\beta \mid u,v,F] \leq e^{-\frac{\sigma^2}{2}\delta_\beta}$. Note that $\delta_\beta \geq (n-2)\cdot\frac{p_{\beta_*}-p'}{1-p'} \geq (n-2)\cdot 2s \geq 24\ln n$. We now set $\sigma = 1/2$ so $c_2 = 1-\sigma = 1/2$. By taking expectation with respect to u and v, we have

$$\mathrm{P}[\alpha \geq c_1\delta_\alpha \mid F] + \mathrm{P}[\beta \leq c_2\delta_\beta \mid F] \leq 2n^{-4}. \tag{12}$$

Since $\tau > 0$, we have that

$$\mathrm{P}[\frac{\alpha}{\beta}I_F \geq \tau] \leq \mathrm{P}[\frac{\alpha}{\beta} \geq \tau \mid (\alpha < c_1\delta_\alpha) \wedge (\beta > c_2\delta_\beta) \wedge F]\ \mathrm{P}[\{(\alpha \geq c_1\delta_\alpha) \vee (\beta \leq c_2\delta_\beta)\} \wedge F]. \tag{13}$$

Under our assumption that $\frac{c_1\delta_\alpha}{c_2\delta_\beta}I_F \leq \tau$ a.s., if $\alpha < c_1\delta_\alpha$, $\beta > c_2\delta_\beta$ and $d_X(u,v) > r$, then $\frac{\alpha}{\beta} < \frac{c_1\delta_\alpha}{c_2\delta_B} \leq \tau$. Therefore, the first term on the right side of (13) is $\mathrm{P}[\frac{\alpha}{\beta} \geq \tau \mid (\alpha < c_1\delta_\alpha) \wedge (\beta > c_2\delta_\beta) \wedge F] = 0$. It then follows from (12) that:

$$\mathrm{P}[\frac{\alpha}{\beta}I_F \geq \tau] \leq \mathrm{P}[(\alpha \geq c_1\delta_\alpha) \vee (\beta \leq c_2\delta_\beta) \mid F] \leq 2n^{-4}$$

Since $\rho_{u,v}(\widehat{G}) \leq \frac{\alpha}{\beta}$, we have $\mathrm{P}[\rho_{u,v}(\widehat{G})I_F \geq \tau] \leq \mathrm{P}[\frac{\alpha}{\beta}I_F \geq \tau] \leq 2n^{-4}$. By union bound, the probability that $\rho_{u,v}(\widehat{G}) < \tau$ for all pairs of nodes $u, v \in V$ satisfying the required conditions is thus at least $1 - \frac{1}{2}n^2(2n^{-4}) = 1 - n^{-2}$.

Finally, for the above argument to hold, we need the assumption $\frac{c_1\delta_\alpha}{c_2\delta_\beta}I_F \leq \tau$ to be satisfied uniformly for all u and v. This comes from the choices and conditions of our parameters; see [19] for details. The lemma then follows. ◀

The above result implies that after Jaccard filtering, although there still may be some extra edges remaining in $\tilde{G}_\tau$, each such edge (u,v) is not really-bad. In fact, $N_{G^*}(u) \cap N_{G^*}(v) \neq \emptyset$ for each such extra remaining edge (u,v), implying that $d_{G^*}(u,v) \leq 2$. This, combined with Lemma 10, essentially leads to the following result. To simplify our statement, we assume $sn = \omega(\ln n)$ in the following result; a more complicated form can be obtained without this assumption (similar to the statement in Lemma 11).

▶ **Theorem 12** (Random Insertion). *Let V be a set of n points sampled i.i.d. from an L-doubling measure $\mu : X \to \mathbb{R}^+$ supported on a compact metric space (X, d_X). Let G^* be the resulting r-neighborhood graph for V; and $\widehat{G}$ a graph obtained by inserting each edge not in G^* independently with probability q. Let $\tilde{G}_\tau$ be the graph after τ-Jaccard filtering of $\widehat{G}$. Then, if Assumption-R holds, $q \leq cs$ and $sn = \omega(\ln n)$, then for $\forall \frac{1}{(6+\frac{1}{\ln n}12c)L^2} \geq \tau \geq (c+2)q + o(1)$, with high probability the shortest path distance metric $d_{\tilde{G}_\tau}$ satisfies: $\frac{1}{2}d_{G^*} \leq d_{\tilde{G}_\tau} \leq d_{G^*}$; that is, $d_{\tilde{G}_\tau}$ is a 2-approximation for d_{G^*} with high probability.*

Proof. Define $\mathcal{E}_1$ to be the event when all the edges in G^* are present in $\tilde{G}_\tau$. By Lemma 10, event $\mathcal{E}_1$ happens with probability at least $1 - n^{-2/3}$. Hence with at least this probability, $d_{\tilde{G}_\tau} \leq d_{G^*}$. We now prove the lower bound for $d_{\tilde{G}_\tau}$.

Let $\mathcal{E}_2$ be the event where for all edges $(u,v) \in E(\tilde{G}_\tau) \setminus E(G^*)$, (u,v) is not really-bad. Lemma 11 says that event $\mathcal{E}_2$ happens with probability at least $1 - n^{-2}$. To this end, observe that if an edge (u,v) is not really-bad, then we have that $d_{G^*}(u,v) \leq 2$ as $N_{G^*}(u) \cap N_{G^*}(v) \neq \emptyset$; specifically, there is a path $u \to w \to v$ connecting u and v through some $w \in N_{G^*}(u) \cap N_{G^*}(v)$.

In what follows, assume both events $\mathcal{E}_1$ and $\mathcal{E}_2$ happen – as discussed above, this assumption holds with high probability due to Lemmas 10 and 11.

Now consider two points $u, v \in V$. First, suppose that u, v are connected in $\tilde{G}_\tau$. Let $\pi = \langle u_0 = u, u_1, \ldots, u_s = v\rangle$ be a shortest path between them in $\tilde{G}_\tau$. Consider each edge (u_i, u_{i+1}) in the shortest path π in $\tilde{G}_\tau$. Either $(u_i, u_{i+1}) \in E(G^*)$, in which case we set $\hat{\pi}(u_i, u_{i+1}) = (u_i, u_{i+1})$. Otherwise if $(u_i, u_{i+1}) \notin E(G^*)$, then (u_i, u_{i+1}) is not really-bad due to event $\mathcal{E}_2$, meaning that $d_{G^*}(u_i, u_{i+1}) \leq 2$. Hence we can find a path $\hat{\pi}(u_i, u_{i+1}) \subset G^*$ of length at most two to connect u_i and u_{i+1} in G^*. Putting these two together, we can construct a path $\hat{\pi} = \hat{\pi}(u_0, u_1) \circ \hat{\pi}(u_1, u_2) \circ \cdots \circ \hat{\pi}(u_{s-1}, u_s)$ connecting $u = u_0$ to $v = u_s$ in G^*. Clearly, this path has length at most $2s$. Hence, for any $u, v \in V$, we have that $d_{G^*}(u,v) \leq 2 d_{\tilde{G}_\tau}(u,v)$ if (u,v) is connected in $\tilde{G}_\tau$.

If u and v are not connected in $\tilde{G}_\tau$, then they are not connected in G^* either; because if there is a path connecting them in G^*, then the same path is present in $\tilde{G}_\tau$ as event $\mathcal{E}_1$ holds. Putting everything together, we then have that with high probability, for any $u, v \in V$, $d_{G^*}(u,v) \leq 2 d_{\tilde{G}_\tau}(u,v)$; that is $d_{\tilde{G}_\tau} \geq \frac{1}{2} d_{G^*}$. The theorem then follows. ◀

4 Combined case

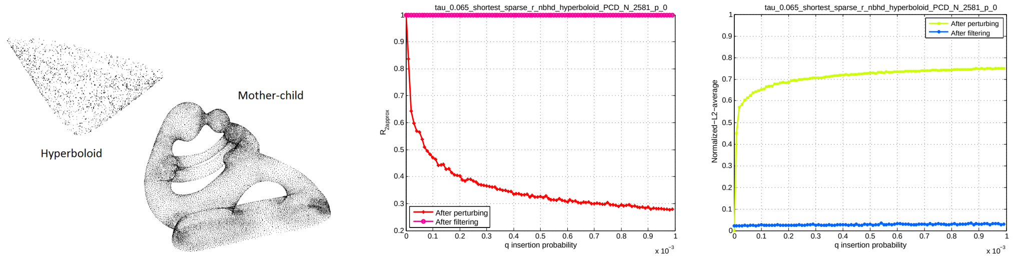
The arguments used in Sections 3.1 and 3.2 can be modified to prove our main result when the observed graph $\widehat{G} = G(r,p,q)$ is generated via the network model described in Definition 3 that includes both edge deletion and insertion. The proof can be found in [19].

▶ **Theorem 13.** *Let V be a set of n points sampled i.i.d. from an L-doubling measure $\mu : X \to \mathbb{R}^+$ supported on a compact metric space (X, d_X). Let G^* be the resulting r-neighborhood graph for V; and $\widehat{G}$ a graph obtained by the network model $G(r,p,q)$ described in Definition 3. Let $\tilde{G}_\tau$ be the graph after τ-Jaccard filtering of $\widehat{G}$. Then, if* Assumption-R *holds, $p \leq \frac{1}{4}$, $q \leq \min\{\frac{1}{8}, cs\}$ and $sn = \omega(\ln n)$, then for any τ such that $\frac{(1-p)^2}{(10 + \frac{5}{3\ln n} + 20c)L^2} \geq \tau \geq \frac{(c+2)q}{1-p} + o(1)$, with high probability the shortest path distance metric $d_{\tilde{G}_\tau}$ is a 2-approximation of the shortest path metric d_{G^*} of the true graph G^*.*

Extension to local doubling measure. We can relax the L-doubling condition of the measure μ where points are sampled from to a *local doubling* condition, where the L-doubling property is only required to hold for metric balls of small radius. Specifically,

▶ **Definition 14** ((R_0, L_{R_0})-doubling measure)**.** Given a metric space $\mathcal{X} = (X, d_X)$, a measure μ on $\mathcal{X}$ is said to be (R_0, L_{R_0})-*doubling* if balls have finite and positive measure and there is a constant L_{R_0} s.t. for all $x \in X$ and any $0 < R \leq R_0$, we have $\mu(\mathrm{B}(x, 2R)) \leq L_{R_0} \cdot \mu(\mathrm{B}(x, R))$.

All our results hold for (R_0, L_{R_0})-doubling measure, as long as the parameter r generating the true graph ${G^*}_r$ satisfies $r < R_0$. The proofs follow the same argument as those for L-doubling measure almost verbatim, and thus are omitted.



(a) (b) (c)

Figure 2 (a) $2.5K$ points sampled from a hyperboloid surface and $24K$ points sampled from mother-child model. (b) Comparison of 2-approximation rate $R_{2approx}$ as insertion probability (x-axis) increases. Top curve is after Jaccard-filtering, while bottom one is for perturbed graph without filtering. (c) Normalized L_2-average error with top curve being the one without filtering, and the bottom one (with significantly lower error) for after Jaccard-filtering. These plots are for hyperboloid case.

5 Some empirical results

We provide some proof-of-principle results to show the effectiveness of the Jaccard filtering process. See [19] for a complete version. There are two sets of experiments.

Synthetic datasets with ground truth. In this experiment we seek to demonstrate that the Jaccard filtering approach works in a robust manner as predicted by our theoretical results. In particular, we start with the following two measures: $\mu_1 : S_1 \to \mathbb{R}^+$ is the "quasi-uniform" measure on the hyperboloid S_1 specified by $x^2 + y^2 - z^2 = 1$ [2]; and $\mu_2 : S_2 \to \mathbb{R}^+$ is a non-uniform measure on the mother-and-child geometric model S_2, where the measure is proportional to the local feature size at each point. For each μ_i, we sample n points V i.i.d and build an r-neighborhood graph (we will specify choice of r later). See Figure 2 (a) for illustration of input samples. This gives rise to a ground-truth neighborhood graph G^*_r. We next generate a set of observed graph $G_{p,q}$, varying the deletion probability (p) and insertion probability (q). Using a fixed parameter τ, we perform τ-Jaccard filtering for each $G_{p,q}$ to obtain a filtered graph $\widehat{G}^{\tau}_{p,q}$. To measure the difference between two metrics D and D', we use two types of error to be introduced shortly. But first, note that since we delete edges, the connectivity of the graph may change. Assume that $D_{i,j} = \infty$ if the two corresponding points p_i and p_j are not connected in the graph. Note that if $D_{i,j} = \infty$ and $D'_{i,j} = \infty$, the relationship $\frac{1}{2}D_{i,j} \leq D'_{i,j} \leq 2D_{i,j}$ still hold.

- **2-approximation rate $R_{2approx}$** is defined by

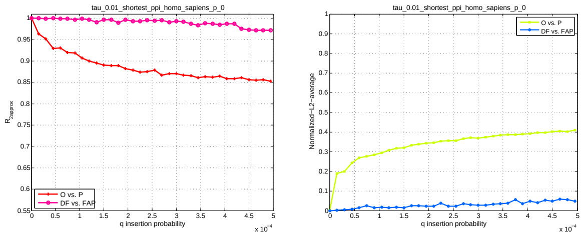
$$R_{2approx}(D, D') = \frac{|\{(i,j), 1 \leq i < j \leq n \mid \frac{1}{2}D_{i,j} \leq D'_{i,j} \leq 2D_{i,j}\}|}{n(n-1)/2}.$$

  In other words, $R_{2approx}$ is the ratio of "good" pairwise distances from D' that 2-approximate those in D.

We also consider L_2 type error. To avoid the cases that $D_{i,j}$ is not comparable with $D'_{i,j}$, we collect the following *good-index set*

$I_{good}(D, D') = \{(i,j), 1 \leq i < j \leq n \mid \text{either } (D_{i,j} < \infty) \wedge (D'_{i,j} < \infty); \text{ or } (D_{i,j} = \infty) \wedge (D'_{i,j} = \infty)\}$.

- **Normalized L_2-average error $\delta_N(D, D')$** is intuitively the root-mean-squared (RMS) error $\delta(D, D')$ normalized by the normalized L_2-norm of D. More specifically,



$R_{2approx}$ normalized L_2

Figure 3 "O vs. P" is the error rate between D_G and D_{G_q}; while "DP vs. FAP" is between D_{G^τ} and $D_{G_q^\tau}$.

$$\delta(D, D') = \sqrt{\frac{\sum_{(i,j)\in I_{good}} (D_{i,j} - D'_{i,j})^2}{|I_{good}|}}; \; \delta_N(D, D') = \frac{\delta(D, D')}{\sqrt{\frac{1}{|\{i<j, D_{i,j}<\infty\}|} \sum_{i<j, D_{i,j}<\infty} D_{i,j}^2}}.$$

Let D_G denote the shortest path metric induced by a graph G. In Figure 2 (b), we compare the 2-approximation rate $R_{2approx}(D_{G^*}, D_{G_q})$ for the sequence of observed graphs G_q for increasing insertion probability q, with $R_{2approx}(D_{G^*}, D_{\widehat{G}_q^\tau})$s for the sequence of filtered graph $\widehat{G}_q^\tau$; while the comparison of the normalized L_2 error $\delta_N(D_{G^*}, D_{G_q})$ versus $\delta_N(D_{G^*}, D_{\widehat{G}_q^\tau})$s for increasing qs is shown in Figure 2 (c). These plots are for the hyperboloid model; those for mother-child model are in [19]. The deletion probability is fixed at $p = 0$, as our experiments show (also matching our theoretical results) that the shortest path metric is rather stable against deletion for a large range of deletion probability. As we can see, randomly inserting edges distorts the shortest path metrics (with low 2-approximation rate and high normalized L_2 error for G_qs). However, our Jaccard-index filtering process restores the metric not only w.r.t 2-approximation rate (which is predicted by our theoretical results), but also w.r.t normalized L_2 error. In this experiment, we choose r (to build the r-neighborhood graph) to be twice of the average distance from a point to its 10-th nearest neighbor in P. The resulting graph for hyperboloid has about 2.5K nodes and 38K edges. Examples where the graphs are much denser are given in [19].

Real network without ground truth. For a given real network G, we can consider it as an observed graph. However, we do not know how this network is generated and there is no ground truth graph G^*. Nevertheless, we carry out the following experiments to indirectly infer the effectiveness of Jaccard-filtering.

Specifically, given G, we gradually add random $(p = 0, q)$-perturbation to it, and compare the shortest path metric D_{G_q} of the perturbed graph G_q with the metric D_G of input network G; q is the insertion probability. Next, we perform τ-Jaccard filtering for all these graphs G and G_qs to obtain G^τ and G_q^τ respectively, and then compare the shortest path metric $D_{G_q^\tau}$ for filtered graphs G_q^τ with D_{G^τ} of G^τ. See Figure 3, where the input is a protein-protein interaction network [14] (6327 nodes and 147547 edges). The distance metric becomes more stable after Jaccard-filtering. More discussions and experiments are in [19], including an example of co-authorships network [18] where Jaccard-filtering shows even bigger improvement.

6 Concluding remarks

Our paper represents one step towards unraveling the structure of the space where data are sampled from. There are many interesting problems along this direction, including how to generalize our network model to better model real networks. We describe one direction here: Our current work recovers the shortest path metric of the hidden graph G. However, there are other common metrics induced from G, such as the diffusion distance metric. In fact, for dense random graphs, say graphs generated from a graphon [10] (including stochastic block models), the spectral structure of such random graphs are stable. This may imply that diffusion distances could also be stable against random perturbations even without any filtering process. Note that such graphs have $\Theta(n^2)$ number edges asymptotically. However, for sparse graphs (which our model could generate), empirically we observe that diffusion distances are not stable under random perturbations. It would be interesting to see whether the Jaccard filtering process (or other filtering procedure) could recover diffusion distances with theoretical guarantees. (Interestingly, we have observed that empirically, Jaccard filtering can recover diffusion distance as well in our experiments.) Finally, it would be interesting to explore whether the analysis and ideas for network models from our paper could be used to create a practical wormhole detector in wireless networks, akin to Ban et al's local connectivity tests based on $[\alpha, \beta]$-rings [3].

Acknowledgement. The authors thank Samory Kpotufe for the pointer to the local version of L-doubling measure.

References

1 Morteza Alamgir and Ulrike V. Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *29th Intl. Conf. Machine Learning (ICML)*, pages 1031–1038, 2012.

2 D. Asta and C. Shalizi. Geometric network comparisons. In *31st Annu. Conf. Uncertainty in AI (UAI)*, 2015.

3 Xiaomeng Ban, Rik Sarkar, and Jie Gao. Local connectivity tests to identify wormholes in wireless networks. In *12th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc'11, pages 13:1–13:11. ACM, 2011.

4 HS Bhadauria and ML Dewal. Efficient denoising technique for CT images to enhance brain hemorrhage segmentation. *Journal of digital imaging*, 25(6):782–791, 2012.

5 Bela Bollobás and Fan R. K. Chung. The diameter of a cycle plus a random matching. *SIAM Journal on discrete mathematics*, 1(3):328–333, 1988.

6 Béla Bollobás and Oliver M. Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.

7 Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2011.

8 Frédéric Chazal, Leonidas J Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41, 2013.

9 Rick Durrett. *Random Graph Dynamics*, volume 20. Cambridge University Press, 2006.

10 Justin Eldridge, Mikhail Belkin, and Yusu Wang. Graphons, mergeons, and so on! In *Advances in Neural Information Processing Systems*, pages 2307–2315, 2016.

11 Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Comm. Review*, 29(4):251–262, 1999.

12 Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, 100(8):4372–4376, 2003.

13 Juha Heinonen. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2012.

14 G. Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, G. R. Gopinath, G. R. Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

15 Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proc. 32nd. ACM Symp. Theory Computing*, pages 163–170. ACM, 2000.

16 Jon Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS)*, pages 431–438. 2002.

17 Elizabeth A. Leicht, Petter Holme, and Mark E. J. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

18 Tiancheng Lou and Jie Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*, 2013.

19 S. Parthasarathy, D. Sivakoff, M. Tian, and Y. Wang. A quest to unravel the metric structure behind perturbed networks. *ArXiv e-prints*, March 2017. `arXiv:1703.05475`.

20 Mathew Penrose. *Random geometric graphs*. Oxford University Press, 2003.

21 Venu Satuluri, Srinivasan Parthasarathy, and Yiye Ruan. Local graph sparsification for scalable clustering. In *ACM SIGMOD Intl. Conf. Management Data*, pages 721–732, 2011.

22 A. Singer and H.-T. Wu. Two-dimensional tomography from noisy projections taken at unknown random directions. *SIAM journal on imaging sciences*, 6(1):136–175, 2013.

23 H. F. Song and X.-J. Wang. Simple, distance-dependent formulation of the Watts-Strogatz model for directed and undirected small-world networks. *Phys. Rev. E*, 90:062801, 2014.

24 Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20, 2003.

25 Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296(5571):1302–1305, 2002. `doi:10.1126/science.1070120`.

26 Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.