Unravel the geometry and topology behind noisy networks

Minghao Tian

Abstract

Graphs and network data are ubiquitous across a wide spectrum of scientific and application domains. An interesting and important observation that arises very often in real networks is the so-called *small-world effect*, which is often referred to as the "six degrees of separation" in popular culture. It basically says that the average shortest-path distance between vertex pairs is very short. Many beautiful generative models for graphs have been proposed, partly aiming to understand this small-world effect observed from real networks. Inspired by the celebrated small-world network model proposed by Watts and Strogatz, a variety of follow-up work consider the model 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). In this proposal, we advocate the perspective that an observed graph is often a noisy version of some discretized 1-skeleton of a hidden domain.

We aim to anaylze two aspects of this type of model — geometry and topology. Specifically, the geometric problem we aim to solve is to recover the metric structure of the hidden domain from the observed graph, which is orthogonal to the usual studies of network models (which often focuses on characterizing / predicting behaviors and properties of real-world networks). We will consider the following natural network model (called *ER-perturbed random geometric graphs* or *noisy random geometric graphs*): We assume that there is a true graph $G^*_{\mathcal{X}}$ which is a certain proximity graph for points sampled from a hidden domain $\mathcal{X}$; while the observed graph $\widehat{G}_{\mathcal{X}}$ is an Erdős–Rényi type perturbed version of $G^*_{\mathcal{X}}$. Two methods are proposed in this research proposal to recover the metric structure of $\mathcal{X}$ from $\widehat{G}_{\mathcal{X}}$: Jaccard-filtering process, which based on Jaccard (similarity) index, and clique-filtering process, which based on *edge clique number* (a local version of the clique number). We show that these two simple filtering processes can recover this metric within a constant multiplicative factor under our network model.

We also consider global and local topological features of the observed graph. We first focus on the clique number of the ER-perturbed random geometric graph $\widehat{G}_{\mathcal{X}}$, which is an important global graph quantity in both network analysis and graph theory. We provide asymptotic tight bounds of the clique number of $\widehat{G}_{\mathcal{X}}$ under different common settings. Then, we take a refined view of the noisy graphs. Specifically, we focus on two types of local subgraphs — *neighborhood subgraphs* $G^{loc}_{u,v}$ for any edge (u, v) in G, which defined as the induced subgraph over the common neighbors of u and v in G, and *rooted (k-neighborhood) subgraphs* G^k_u for any vertex u in G, which is the induced subgraph over the vertices within k distance (shortest-path distance) away from u. We show that the edge clique number in $\widehat{G}_{\mathcal{X}}$ presents two fundamentally different types of behaviors, depending on which "type" of randomness it is generated from. Also, we notice that many graph representations proposed recently are based on rooted subgraphs (or similar substrutures), which may be later used in tasks like network comparison and network classification. However, as we know, the theoretical understanding of the topology of these subgraphs is rather limited. We take a first step to explore the topological features of the rooted subgraphs in Erdős–Rényi random graph $G(n, p)$. Specifically, we show that the *1-dimensional Betti number* of 1-*ring subgraphs* (the induced subgraph over vertices exactly 1 distance away from the randomly selected root vertex) satisfies a central limit theorem.

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 graph-based data mining.

Among them a varity of empirical studies focus on the graph properties (such as the degree distribution and the clustering coefficient) of different types of real networks. An interesting and important observation that arises very often and has practical implications is the celebrated *small-world effect* [41] discovered in a seminal work of Milgram [37], which is often referred to as the "six degrees of separation" in popular culture. It basically says that the average shortest-path distance, appropriately defined, between vertex pairs is very short. Although first studied in friendship networks, this phenomenon appears to be occurring in almost all types of networks.

Many beautiful generative models for graphs have been proposed, partly aiming to understand this small-world effect observed from real networks [41, 56]. One of the most classic models with theoretical guarantee is the *Erdős–Rényi random graph model* $G(n, p)$ [16, 17], constructed by adding edges between all pairs of n vertices with probability p independently. However, since this model is purely combinatorial, it fails to capture the geometry (shape) of the network. For example, most people make friends based on common interests, location, age and so on. In other words, in friendship networks, vertices could be sampled from some feature space of people, and two people could be connected if they are nearby in the feature space. Obviously, the structure of the feature space cannot be encoded in the Erdős–Rényi random graph model $G(n, p)$.

Another 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 model introduced by Watts and Strogatz [56] generates a network by starting with a k-nearest neighbor graph spanned by vertices 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 displays two important characteristics seen in small-world 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 small-world networks, such as adding random edges in a distance-dependent manner [28, 50], or extending similar ideas to incorporate hierarchical structures in networks; e.g, [29, 55]. See [30] for a survey on this topic.

Statement of the problems. Inspired by the small-world model by Watts and Strogatz (and some later variants), we take the perspective that an observed graph can be deemed as a noisy snapshot of the graph representation (discretized 1-skeleton) of a hidden domain of interest. However, orthogonal to the usual studies of this type of network models (which often focuses on simulating real-world networks and interpreting the observed phenomena), we aim to answer the following two main questions:

1. (Geometry) *What can we infer about the hidden domain from the observed graph?*

2. (Topology) *What are the topological properties (such as the clique number) of the observed graph?*

To be more specific, we propose the following network model in [45]: Assume that the hidden space that generates data is a "nice" probability measure μ supported on a compact metric space $\mathcal{X} = (X, d_X)$ (e.g, the uniform probability measure supported on an embedded smooth low-dimensional Riemannian manifold). Suppose that the data points V are sampled i.i.d from this probability 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 $\widehat{G}$ 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 vertices u, v could be inserted to G_r^* with probability q. This model is later called the *ER-perturbed random geometric graph* [25]

or the *noisy random geometric graph* [26], as the true proximity graph G_r^* generated in this way is in fact a *random geometric graph* in random graph theory [46].

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 vertices) 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 $\widehat{G}$ as an ER-type perturbation of the proximity graph G_r^* to account for such exceptions.

Assumptions and notations. All the graphs mentioned in this proposal are simple undirected graphs, which means there are no duplicate edges or loops (a loop is an edge that connects a vertex to itself). For any graph G, let $V(G)$ and $E(G)$ refer to its vertex set and edge set, and let $N_G(u)$ denote the set of neighbors of vertex u in G (i.e. vertices connected to $u \in V(G)$ by edges in $E(G)$). We also use the terminology *with high probability*. If $A_1, A_2, \cdots$ is a sequence of events, then "A_n happens with high probability" means that $\lim_{n\to\infty} \mathbb{P}[A_n] = 1 + o(1)$.

1.1 Geometry — Metric recovery

It is proved (Theorem 2.5 in [45]) that under some general assumptions ($\mathcal{X}$ is a compact geodesic metric space equipped with a doubling measure μ), for any two vertices in the hidden metric space $\mathcal{X}$, the distance between them in $\mathcal{X}$ can be approximated by rescaling the shortest-path distance between them in the induced $r-$neighborhood graph G_r^*. Thus, the main geometric problem we aim to solve is to recover the metric structure of G_r^*, which reflects that of the hidden space $\mathcal{X}$, from the observed graph $\widehat{G}$.

There are different motivations for this problem. 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 $\widehat{G}$ to get a more accurate representation of G_r^*. Indeed, in [18], Godberg and Roth empirically show how to use small-world model to help remove false edges in protein-protein interaction (PPI) networks. See [5] for more examples.

1.2 Topology — Graph properties

The observed graph $\widehat{G}$ generated by adding p-deletion and $q-$insertion (ER type perturbation) to a random geometric graph G_r^* is of interest itself. $\widehat{G}$ in some sense is a mixed model of two classic random graph models — Erdős–Rényi random graph and random geometric graph. The main topological problem we aim to solve is to get a topological summary of the observed graph $\widehat{G}$ based on its graph properties (such as the clique number).

In fact, $\widehat{G}$ is related to the continuum percolation theory [36]. However, our understanding about the model so far is still limited: In previous studies, the underlying spaces are typically plane (called the Gilbert disc model) [8], cubes [11] and tori [23]; the vertices are often chosen as the standard lattices of the space; and the results usually concern the connectivity [9, 47] or diameter (e.g, [57]).

To be more concrete, we consider the following two problems.

1. Global property — the clique number
Cliques in graphs are important objects in many application domains (e.g, in social networks [20], chemistry [13] and PPI networks [51]). Not surprisingly, the occurrence of a clique is often viewed as a highly cohesive subgroup in social networks [41]. The clique number of a graph is the number of vertices in a maximum clique in the graph. In other words, it is the largest dimension of the non-trivial homology group of the induced clique (flag) complex, which is now a popular topic in network analysis [52, 27, 22] due to the fast developing field called *topological data analysis* (TDA).

We are interested in the clique number of the observed graph $\widehat{G}$. From the theoretical aspect, the clique number in Erdős–Rényi random graph has been studied extensively in 20th century [19, 7], and is a relative new topic in random geometric graph [46]. Even in the case when we only add q-insertion, which means that the observed graph $\widehat{G}$ can be viewed as a union of these two types of random graphs, understanding the behavior of the clique number of each of them seems not enough. In general, the clique number of the union graph $G = G_1 \cup G_2$ of two graphs G_1 and G_2 could be significantly larger than the clique number in each individual graph G_i: Consider for example G_1 is a collection of $\sqrt{n}$ disjoint cliques, each of size $\sqrt{n}$, while G_2 equals to the complement of G_1. The union $G_1 \cup G_2$ is the complete graph and the clique number is n. However, the clique number of G_1 or of G_2 is $\sqrt{n}$.

2. Properties of local subgraphs

Besides the global topological properties like the clique number, inspired by the network motifs analysis [38], we also consider the topological properties of local subgraphs of our model $\widehat{G}$. Naturally, for any graph $G = (V, E)$, there are two types of local subgraphs:

1. *Neighborhood subgraphs*: defined for each edge $(u, v) \in E$ as the induced subgraph over the set of vertices $(N_u \cap N_v) \cup \{u\} \cup \{v\}$. Denote such subgraph by $G^{loc}_{u,v}$;
2. *Rooted (k-neighborhood) subgraphs*: for a given integer k, defined for each vertex $u \in V$ as the induced subgraph rooted at u over the set of vertices $\Gamma^k_u = \{v \in V : d_G(u, v) \leq k\}$, where d_G is the shortest-path metric. Denote such subgraph as G^k_u.

We remark that both of the subgraphs are used intensively in network analysis and data mining, not only to understand and further predict the individual behavior [59, 2] but also to find representations of the global graph by using features observed in these subgraphs [40, 12, 42, 58].

We consider the topological properties for both types of subgraphs. For neighborhood subgraphs, we introduce a local adapted version of the clique number called "*edge clique number*" [25, 26], which is defined for each edge (u, v) in a graph G as the clique number of $G^{loc}_{u,v}$ (or equivalently, the size of the largest clique containing (u, v) in G). Specifically, we consider the edge clique numbers of the observed graph $\widehat{G}$ (noisy random geometric graph). Roughly speaking, there are two types of edges in $\widehat{G}$: "long-range" edges generated from q-insertion and "short" edges generated from the geometric graph. We are interested in comparing the edge clique numbers of both types of edges.

To study rooted (k-neighborhood) subgraphs, inspired by the work on the topolgy of metric graphs [15, 44], we consider the *extended persistence diagram* (induced by the distance-to-root function) for each rooted subgraph (rooted at any vertex) of an observed graph. A topological representation of the graph can be derived by combining all these diagrams. There exist a variety of approaches proposed to get different representations of graphs and further used for graph comparison (e.g., see [43, 31] for surveys on graph kernels). More recently, methods using neural networks are developed [32, 60, 39]. Although empirically they show competitive performance, the theoretical understanding is rather limited. We take a first step to explore the topological features of the rooted subgraphs in Erdős–Rényi random graph $G(n, p)$ from the perspective of random graph theory.

2 Current progress

In this section, all the problems mentioned in Section 2 will be discussed. Before answering the questions, we first give a detailed rigorous mathematical description of our graph model. Suppose we are given a compact geodesic metric space $\mathcal{X} = (X, d)$ [10] [1]. We consider "nice" measures on $\mathcal{X}$. Specifically,

[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. Uniqueness of geodesics is not required. Riemannian manifolds or path-connected compact sets in the Euclidean space are all geodesic metric spaces.

Definition 1 (Doubling measure [21])**.** *Given a metric space $\mathcal{X} = (X, d)$, let $B_r(x) \subset X$ denotes the closed metric ball $B_r(x) = \{y \in X \mid d(x, y) \leq r\}$. A measure μ on $\mathcal{X}$ is said to be* doubling *if every metric ball (with positive radius) has finite and positive measure and there is a constant $L = L(\mu)$ s.t. for all $x \in X$ and every $r > 0$, we have $\mu(B_{2r}(x)) \leq L \cdot \mu(B_r(x))$. We call L the* doubling constant *and say μ is an L-*doubling measure*.*

Intuitively, the doubling measure generalizes a nice measure on the Euclidean space, but still behaves nicely in the sense that the growth of the mass within a metric ball is bounded as the radius of the ball increases.

ER-perturbed random geometric graph [45]. We consider the following random graph model: Given a compact metric space $\mathcal{X} = (X, d)$ and a L-doubling probability measure μ supported on X, let V be a set of n points sampled i.i.d. from μ. We build the $r-$neighborhood graph $G_r^*(\mathcal{X}) = (V, E^*)$ for some parameter $r > 0$ on V; that is, $E^* = \{(u, v) \mid d(u, v) \leq r, u, v \in V\}$. $G_r^*(\mathcal{X})$ is often called a *random geometric graph* [46] generated from $(\mathcal{X}, \mu, r)$. Now we add the following two types of random perturbations:

p-deletion: For each existing edge $(u, v) \in E^*$, we delete edge (u, v) with probability p.

q-insertion: For each non-existent edge $(u, v) \notin E^*$, we insert edge (u, v) with probability q.

The order of applying the above two types of perturbations doesn't matter since they are applied to two disjoint sets respectively. The final graph $\widehat{G}_r^{p,q}(\mathcal{X}) = (V, \hat{E})$ is called a *(p, q)-perturbation of $G_r^*(\mathcal{X})$*, or simply an *ER-perturbed random geometric graph*. The reference $\mathcal{X}$ and parameters r, p, q are sometimes omitted from the notations when their choices are clear.

In what follows, we will first focus on geometry — the metric recovery problem, where two denoising methods with theoretical guarantee are presented. Later we will move to the study of topological features of the local subgraphs. For clarity and conciseness, some results are given in a very brief manner and all the proofs are omitted. The complete statements of the theorems and their proofs can be found in the corresponding cited papers.

2.1 Geometry

First recall the metric recovery problem introduced in Section 2.1:

Shortest-path metric recovery
Input: An observed graph $\widehat{G} = \widehat{G}_r^{p,q}(\mathcal{X})$
Output: Recover (approximately) the shortest path metric $d_{G_r^*(\mathcal{X})}$ from $\widehat{G}$

To describe the proximity, we define the so-called c-approximation as follows.

Definition 2 (c-approximation)**.** *Let G and G' be two graphs on the same set of nodes V, and equipped with graph shortest path metric d_G and $d_{G'}$, respectively. By $d_G \leq cd_{G'}$, we mean that for any two nodes $u, v \in V$, we have that $d_G(u, v) \leq cd_{G'}(u, v)$. We say that $d_{G'}$ is* a c-approximation of d_G *if $(1/c)d_G \leq d_{G'} \leq cd_G$.*

Suppose for any two nodes u, v connected in G_r^* share sufficient number of common neighbors, intuitively, even after removing a constant fraction of edges in G_r^*, we can still guarantee that with high probability u and v will have some common neighbors left, and thus u and v can be connected through that common neighbor by a path of length 2 in $\widehat{G}_r^{p,0}$. In fact, under the following density-condition, we can prove that with high probability, the shortest-path metric $d_{\widehat{G}_r^{p,0}}$ is a 2-approximation of the shortest-path metric $d_{G_r^*}$ for even constant deletion probability p [45].

(Density-cond) The parameter r and the doubling measure μ satisfy the following condition: There exists $\mathrm{s} \geq 13 \ln n/n \, (= \Omega(\ln n/n))$ such that for any $x \in X$, $\mu\left(B_{r/2}(x)\right) \geq \mathrm{s}$.

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 constructed as the r-neighborhood graph for points sampled i.i.d. from a uniform measure on a Euclidean cube, asymptotically this is the same requirement so that the resulting r-neighborhood graph is connected with high probability [46].

However, the long-range edges (say edge $(u, v) \in \widehat{G}_r^{p,q}$ with $d_X(u, v) > 2r$) generated by q-insertion can extensively distort the shortest-path metric of G_r^*. This also reveals the essence why Watts and Strogatz's network model (see Section 1) has the small-world effect. Therefore, the key point of shortest-path metric recovery is to remove those long-range edges. We call this task as *filtering*. In what follows, we give two graph filtering techniques, which can be proved to remove almost all the long-range edges in $\widehat{G}_r^{p,q}$ with high probability, and thus can be further used to construct a filtered graph $\tilde{G}$ such that $d_{\tilde{G}}$ approximates $d_{G_r^*}$.

2.1.1 Filtering by using Jaccard (similarity) index

In this section, we show that a simple filtering process based on the so-called *Jaccard (similarity) index* can be used to recover the shortest-path metric of G_r^* up to a factor of 2 with high probability.

Definition 3 (Jaccard (similarity) index)**.** *Given any edge $(u, v) \in E(G)$ in any graph 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 vertices connected by an edge in a graph [33], and has been commonly used in practice for denoising and sparsification purposes [49, 48]. Our results provide a theoretical understanding for such empirical Jaccard-based denoising approaches.

We now propose the following Jaccard-Index-based filtering process, which we call a τ-Jaccard filtering, as it uses a parameter τ. We represent the output filtered graph as $\tilde{G}_\tau^J$:

τ-Jaccard filtering: Given graph $\widehat{G}$, we construct another graph $\tilde{G}_\tau^J$ on the same vertex set as follows: for each edge $(u, v) \in E\left(\widehat{G}\right)$, we insert the edge (u, v) into $E\left(\tilde{G}_\tau^J\right)$ if and only if $\rho_{u,v}\left(\widehat{G}\right) \geq \tau$. That is, $V\left(\tilde{G}_\tau^J\right) = V\left(\widehat{G}\right)$ and $E\left(\tilde{G}_\tau^J\right) := \left\{(u, v) \in E\left(\widehat{G}\right) \mid \rho_{u,v}(\widehat{G}) \geq \tau\right\}$.

We have the following theorem to show the effectiveness of τ-Jaccard filtering.

Theorem 4 (Metric recovery based on Jaccard filtering [45], simplifed)**.** *Under* Density-cond*, and assume $p \leq \frac{1}{4}, q \leq s$ with $s = \omega(\ln n/n)$, there exists a range of τ such that the shortest-path metric $d_{\tilde{G}_\tau^J}$ 2-approximates $d_{G_r^*}$ with high probability, where $\tilde{G}_\tau^J$ is the filtered graph of $\widehat{G}_r^{p,q}$ by τ-Jaccard filtering.*

Remark. The insertion probaility can be increased at the cost of decreasing the range of valid τ's. Also note that the insertion probability cannot be larger than s which affects the number of points falling in any $r/2$ ball. It is reasonable since once the underlying graph G_r^* is dense enough, the diameter of G_r^* (the largest shortest-path distance between all pairs of vertices) will be small (probably a constant) and thus the q-insertion won't distort the shortest-path metric much. However, the method introduced in the next section shows that even if G_r^* is not extremely dense, or specifically, the number of points falling in any $r/2$ ball

is only $\Theta(\ln n)$, we can still handle very large insertion probability, say $q = o(1)$, due to the geometry of the underlying space. It is quite surprising as it shows that even if the number of edges inserted is of order $\omega(n \ln n)$ (can almost reach $\Theta(n^2)$), which is much larger than the number of edges in G_r^* (which is $\Theta(n \ln n)$), we can still recover the shortest-path metric of G_r^*.

2.1.2 Filtering by using edge clique number

First recall that $G_{u,v}^{loc}$ the neighborhood subgraph at edge (u, v) is the induced subggraph over the set of vertices $(N_u \cap N_v) \cup \{u\} \cup \{v\}$. We now give the defintion of the so-called *edge clique number*.

Definition 5 (Edge clique number [25, 26])**.** *Given a graph $G = (V, E)$, for any edge $(u, v) \in E$, its* edge clique number $\omega_{u,v}(G)$ *is defined as*

$$\omega_{u,v}(G) = \textit{the clique number of the neighborhood subgraph } G_{u,v}^{loc}.$$

Easy to see that it is also the size of the largest clique in G containing edge (u, v). Similar to τ-Jaccard filtering process, we introduce our edge-clique-number based filtering process.

γ-Clique filtering: Given graph $\widehat{G}$, we construct another graph $\tilde{G}_\gamma^C$ on the same vertex set as follows: For each edge $(u, v) \in E\left(\widehat{G}\right)$, we insert the edge (u, v) into $E\left(\tilde{G}_\gamma^C\right)$ if and only if $\omega_{u,v}\left(\widehat{G}\right) \geq \gamma$. That is, $V\left(\tilde{G}_\gamma^C\right) = V\left(\widehat{G}\right)$ and $E\left(\tilde{G}_\gamma^C\right) := \left\{(u, v) \in E\left(\widehat{G}\right) \mid \omega_{u,v}\left(\widehat{G}\right) \geq \gamma\right\}$.

We need the following technical assumption called Assumption-R on the parameter r and probability measure μ, where an additional condition is required along with Density-cond.

[Assumption-R]: The parameter r and the doubling measure μ satisfy the following condition:

There exist $\mathrm{s} \geq \frac{13 \ln n}{n} \left(= \Omega(\frac{\ln n}{n})\right)$ and a constant ρ such that for any $x \in X$

(Density-cond) $\mu\left(B_{r/2}(x)\right) \geq \mathrm{s}$.
(Regularity-cond) $\mu\left(B_{r/2}(x)\right) \leq \rho \mathrm{s}$

It basically says that the mass contained inside all radius-r metric balls are similar (within a constant ρ factor); so the measure μ is roughly uniform at this scale r and thus the number neighbors of any vertex in G_r^* is $\Theta(sn)$ (can potentially be $\Theta(\ln n)$ if pick $s = \Theta(\ln n/n)$). These conditions can be satisfied when the input measure is the so-called (Ahlfors) d-regular measure [21], which is in fact stronger and essentially requires that such a bound on the mass in a metric ball $B_{r'}(x)$ holds for every radius r'.

We now show the result for $\widehat{G}_r^{0,q}$ where only q-insertion is added to G_r^*.

Theorem 6 (Metric recovery based on Clique filtering [25], q-insertion only, simplified)**.** *Under* Assumption-R*, there exist constant c_1, c_2, c_3 such that if $\gamma < sn/4$ and $q \leq \min\left\{ c_1,\ c_2 (1/n)^{c_3/\gamma} (\gamma/\mathrm{sn}) \right\}$, then the shortest-path metric $d_{\tilde{G}_\gamma^C}$ 3-approximates $d_{G_r^*}$ with high probability, where $\tilde{G}_\gamma^C$ is the filtered graph of $\widehat{G}_r^{0,q}$ by γ-Clique filtering.*

Remark. Here we give an example of the above theorem. If we choose $\gamma = \ln n$ and assume that $sn > 4\gamma$, then with high probability we can recover the shortest-path metric within a factor of 3 as long as $q \leq c\frac{\ln n}{sn}$ for some constant $c > 0$. If $sn = \Theta(\ln n)$ (but $sn > 4\gamma = 4 \ln n$), then q is only required to be smaller than a (sufficiently small) constant. Next, if $sn = \ln^a n$ for some $a > 1$, then we require that $q \leq c/(\ln^{a-1} n)$. In constrast, the Jaccard-filtering process (Theorem 4) requires that $q = o(s)$, which is $q = o(\ln^a n/n)$. The gap (ratio) between these two bounds is nearly a factor of n.

We also have the result for $\tilde{G}_r^{p,q}$ (omitted intentionally for conciseness; See Theorem 4.3 in [25] for the complete result). It can be proved that for a *constant* deletion probability p, our clique filtering process still requires a much larger range of insertion probability q compared to what's required by Jaccard filtering, although the gap is much smaller than the case for $p = 0$.

However, we do point out that the Jaccard-filtering process is algorithmically much simpler and faster, and can be done in $O(n^2)$ time, while the clique-filtering requires the computation of edge-clique numbers, which is computationally expensive.

2.2 Topology

In this section, we focus on the topological features. We first consider the clique number of noisy random geometric graph (ER-perturbed random geometric graph generated in Euclidean space). As we mentioned in Section 1, the clique number of random graphs is not only an important graph property [7, 46], but also a central topic in the topological analysis on random graphs, especially in the analysis of random clique (flag) complexes [24, 6]. The reason why we discuss the Euclidean version of noisy random geometric graph is to align with the standard notations and assumptions often seen random graph literature. The conclusions derived below can also be extended for general ER-perturbed random geometric graphs.

In the second part of this section, we discuss the topology of two types of subgraphs — neighborhood subgraphs and rooted subgraphs, both of which are mentioned early in Section 1. The result on edge clique numbers (clique numbers for neighborhood graphs) are given there. Not surperisingly, the effectiveness of Clique-filtering process (Theorem 6) can be derived directly by using those results.

The other object we are interested in is the extended persistence diagram for each rooted subgraph (treated as a metric graph) in Erdős–Rényi random graphs $G(n, p)$. In particular, we study a specific type of points in the diagram, which can be easily interpreted as the *cyclomatic number* (or the 1st *Betti number*) of 1-ring of the rooted subgraph. We show that the number of this type of points satisfies a central limit theorem, which gives some basic understanding of the diagrams generated by local subgraphs.

2.2.1 Global property — the clique number

In this section, to align with the standard defintion of random geometric graph [46], we focus on the following setting of noisy random geometric graphs, which we refer to as *the standard-setting*:

- The space we consider is the d-dimensional Euclidean space $\mathbb{R}^d$ with a fixed dimension d, equipped with some arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$.
- ν is a probability distribution with finite maximum density σ; and $X_1, X_2, \cdots$ are independent random variables sampled from ν.
- $r = (r(1), r(2), \cdots)$ is a sequence of positive real numbers such that $r(n) \to 0$ as $n \to \infty$.
- p and $q = q(n)$ are real numbers between 0 and 1 (for simplicity, we only consider the case when p is a fixed constant).
- $G_n, G_n^{p,q}$ denote the *random geometric graph* $G(X_1, \cdots, X_n; r(n))$ and its (p, q)-*perturbation* (p-deletion, q-insertion), respectively.

Remark. Different from the Density-cond or Assumption-R in Section 2.1, here in standard random geometric graphs, only the upper bound of the density function is required, which means we now only have an upper bound for the number of points in any "local ball" (similar to Regularity-cond in Assumption-R).

We use the terminology *almost surely* (or a.s.): In particular, if $\xi_1, \xi_2, \cdots$ is a sequence of random variables and $k_1, k_2, \cdots$ is a sequence of positive numbers, then $\xi_n = O(k_n)$ *a.s.* means that there exist

$C_1 > 0$ such that $\mathbb{P}[\xi_n \leq C_1 k_n] \to 1$ as $n \to \infty$. Similarly, $\xi_n = \Omega(k_n)$ *a.s.* and $\xi_n = \Theta(k_n)$ *a.s.* are also defined in the same pattern.

Many properties of $G(X_1, \cdots, X_n; r)$ are qualitatively different depending on which distance $r = r(n)$ is chosen. In some sense, the distance r here plays a role similar to the edge-adding probability $p(n)$ in Erdős–Rényi random graphs $G(n, p)$. Following standard settings in the literature [35, 46], we consider the following three regimes of r, or more precisely, of the quantity nr^d:

I. (*"very sparse"*) $nr^d \leq n^{-\alpha}$ for some fixed $\alpha > 0$;

II. (*"quite sparse"*) $n^{-\epsilon} \ll nr^d \ll \ln n$ for all $\epsilon > 0$;

III. (*"dense"*) $\sigma nr^d / \ln n \to t \in (0, \infty)$;

Recall that the clique number of a graph is the number of vertices in a maximum clique in the graph. We denote the clique number of any graph G by $clique(G)$. For simplicity, we only show the result for the case when nr^d is in the dense regime. Results for other regimes can be found in [26].

Theorem 7 (The clique number of noisy random geometric graphs, "dense" regime [26])**.** *Suppose $G_n^{p,q}$ is a (p, q)-perturbed noisy random geometric graph in the standard-setting with a constant $p \in (0, 1)$. Then there exists a constant $T > 0$ such that if $\sigma nr^d / \ln n \to t \in (T, \infty)$, then there exists two constants C_1, C_2 such that*

a) if $q \leq (1/n)^{C_1 / \ln \ln n} (\ln \ln n / \ln n)$, then

$$clique\left(G_n^{p,q}\right) = \Theta\left(\ln\left(nr^d\right)\right) \quad a.s.$$

b) and if $q = \Theta(1)$ and $q \leq C_2$, then

$$clique\left(G_n^{p,q}\right) = \Theta\left(\log_{\frac{1}{q}} n\right) \quad a.s.$$

Remark. The theorem above basically says that under those setting of parameters r, p, q, we can get a tight asymptotic bound for the clique number. However, there is a gap of q between the conditions of a) and b), where only very loose bound can be derived currently. One potential future work is to develop techniques to solve the clique number problem when q falls in this gap.

2.2.2 Properties of local subgraphs

In this section, we consider the topological properties of two types of (local) subgraphs — neighborhood subgraphs and rooted subgraphs (see Section 1). Specifically, we first discuss the behavior of the clique numbers of neighborhood subgraphs (equivalently, the edge clique numbers) induced by different types of edges in $\widehat{G}_r^{p,q}(\mathcal{X})$. After that we give some understanding of the extended persistence diagrams (induced by the *super-level set filtration* based on the the distance-to-root function) of rooted subgraphs in Erdős–Rényi random graphs $G(n, p)$.

1. The behavior of the edge clique numbers in $\widehat{G}_r^{p,q}(\mathcal{X})$

Recall that the edge clique number of any edge (u, v) in any graph G, denoted by $\omega_{u,v}(G)$, is the clique number of the neighborhood subgraph $G_{u,v}^{loc}$. In what follows, we show that the edge clique number in $\widehat{G}_r^{p,q}(\mathcal{X})$ presents two fundamentally different types of behaviors, depending on which "type" of randomness it is generated from. We would like to point out that the result on the effectiveness of Clique-filtering process in Section 2.1.2 is a direct application of this observation. Again, we only give the result for the insertion-only case for simplicity.

Theorem 8 (Two different behaviors of edge clique number [25], q-insertion only, simplified)**.** *Under* Assumption-R*, for any insertion probability $q = o(1)$, with high probability,*

- *for all* "good edge" $(u, v) \in \widehat{G}_r^{0,q}$ *(i.e., $d_X(u, v) \leq r$), we have $\omega_{u,v}\left(\widehat{G}_r^{0,q}\right) \geq sn$;*
- *for all* "bad edge" $(u, v) \in \widehat{G}_r^{0,q}$ *(i.e., $d_X(u, v) \geq 3r$), we have $\omega_{u,v}\left(\widehat{G}_r^{0,q}\right) = o(sn)$;*

Remark. The above theorem shows that there is a gap between the edge clique number of edges generated by the random geometric graph G_r^* ("good edges") and some "very long" edges generated by $(0, q)$-perturbation ("bad edges"), and thus we can further differentiate these two types of edges by using edge clique number. Moreover, this result also provides topological information about the neighborhood subgraphs, which may be potentially used in some representation of the global graph.

2. Topology of rooted subgraphs in $G(n, p)$

We now move to the rooted subgraphs. As we mentioned in Section 1, many graph representations proposed recently are based on rooted subgraphs (or similar substrutures). However, as we know, the theoretical understanding of the topology of these subgraphs is limited. We take a first step to analyze the topology of rooted subgraphs in Erdős–Rényi random graph $G(n, p)$.

Recall that given a positive integer k, a *rooted (k-neighborhood) subgraph* of any graph G at vertex u, denoted by G_u^k, is defined as the induced subgraph rooted at u over the set of vertices $\Gamma_u^k = \{v \in V : d_G(u, v) \leq k\}$, where d_G is the shortest-path metric. By setting different k, we can get different scales of local subgraphs. Furthermore, we define the so-called *l-ring subgraphs* as follows, which provide a refined view of root subgraphs.

Definition 9 (l-ring subgraphs [53])**.** *Given a graph $G = (V, E)$, for any vertex $u \in V$, the* l-ring subgraph *of u is the induced subgraph over the vertex set $\Delta_u^l = \{v \in V : d_G(u, v) = l\}$, where d_G is the shortest-path metric. Denote such subgraph by $G_{l \to u}$.*

For a given rooted subgraph G_u^k, it can be viewed as a metric graph [15] equipped with the distance-to-root function $f_{G_u^k}$: for any vertex v in G_u^k, $f_{G_u^k}(v) = d_{G_u^k}(v, u)$; then we do linear interpolation for each edge. For example, suppose z is the mid point of edge (v, w), then $f_{G_u^k}(z) = \left[f_{G_u^k}(v) + f_{G_u^k}(w)\right]/2$.

Now consider the 1-dimensional extended persistence diagrams [4] of these rooted subgraphs induced by the super-level set filtration of the distance-to-root function. A special type of points in the diagrams is the points on the diagonal. By an argument on reading the *Betti number* of some substructures from the extended persistence diagrams (Theorem 2 in [4]), it is easy to see that the number of (t, t) in the diagram associated with vertex u is the 1-dimensional Betti number of $G_{t \to u}$ (as an 1-dimensional simplicial complex). Equivalently, this quantity can also be interpreted as the so-called *cyclomatic number* of $G_{t \to u}$, which is the minimum number of edges that must be removed from the graph to break all its cycles.

In particular, we are interested in the behavior of the number of $(1, 1)$ points in the extended persistence diagrams of the rooted subgraphs in Erdős–Rényi random graph $G(n, p)$. In other words, we consider the cyclomatic numbers of 1-ring subgraphs of $G(n, p)$. More precisely, we focus on the bahavior of the random variable $C_{n,p}^{(1)}$ defined as follows:

1. Sample a graph G from $G(n, p)$;
2. Randomly pick a vertex v in G;
3. Set $C_{n,p}^{(1)} := \beta_1(G_{1 \to v})$, where $\beta_1(\cdot)$ is the 1-dimensional Betti number;

A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables is said to *converge weakly* to a limiting random variable X (written $X_n \Rightarrow X$) if $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for all bounded continuous function f.

Now we are ready to state our result on $C_{n,p}^{(1)}$.

Theorem 10 (Central limit theorem for $C_{n,p}^{(1)}$ [53])**.** *If $n^{-1/2} \ll p < 1$, then*

$$\frac{C_{n,p}^{(1)} - \mathbb{E}\left[C_{n,p}^{(1)}\right]}{\sqrt{\mathrm{Var}\left[C_{n,p}^{(1)}\right]}} \Rightarrow \mathcal{N}(0,1)$$

Remark. In the above theorem, we only consider the 1-dimensional Betti number (cyclomatic numbers) of 1-ring subgraphs. Based on some empirical results, we conjecture that the 1-dimensional Betti number for l-ring subgraphs with $l \geq 2$ (if exist) also satisfies a central limit theorem in some different range of p. Let $\beta_{1,d}(G) := |\{v \in V : \beta_1(G_{1\to v}) = d\}|$, where $\beta_1(\cdot)$ is the 1-dimensional Betti number. We also conjecture that the distribution of $\beta_{1,d}(G(n,p))$ should follow a Normal distribution. Check Section 3.2.2 for more discussions.

3 Future directions

Here are some directions either I am currently working on or may explore in the future.

3.1 Geometry

Recall the *shortest-path metric recovery* problem mentioned at the beginning of Section 2.1: we want to recover (approximately) $d_{G_r^*(\mathcal{X})}$ the shortest path metric of the random geometric graph from the (p,q)-perturbation of $G_r^*(\mathcal{X})$. We proposed two filtering techniques, τ-Jaccard filtering and γ-Clique filtering, to solve this problem. τ-Jaccard filtering is algorithmically much simpler and faster, but theoretically speaking, it cannot be used to handle much larger noise (say the insertion probability q is large). Although such situation can still be solved by γ-Clique filtering, in general, finding maximum cliques (sub)graphs (or even approximating it) is algorithmically expensive. Thus, we are looking for other graph quantities, which is easy to compute like Jaccard index and can be applied to the noise level where γ-Clique filtering still works.

3.1.1 Filtering by using truss number

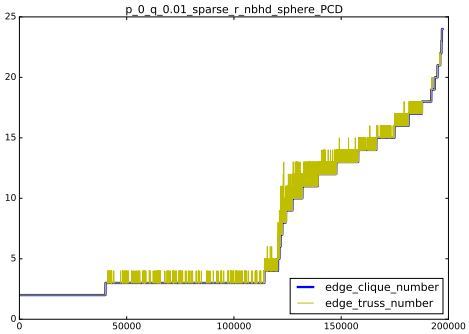
Given a graph G, the k-truss of G is the largest subgraph of G in which every edge is contained in at least $(k-2)$ triangles within the subgraph [54]. We now give the definition of the so-called *edge truss number* (or *trussness*).

Definition 11 (Edge truss number [54])**.** *Given a graph $G = (V, E)$, for any edge $(u, v) \in E$, its* edge truss number $T_{u,v}(G)$ *is defined as*

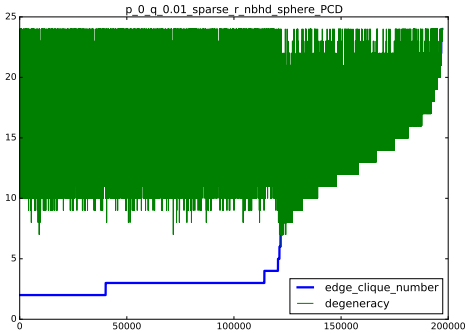
$$T_{u,v}(G) = \max\left\{k : \textit{ the neighborhood subgraph } G_{u,v}^{loc} \textit{ has a } k\textit{-truss}\right\}.$$

For example, if $N_u(G) \cap N_v(G) = \emptyset$, then $T_{u,v}(G) = 2$; if $|N_u(G) \cap N_v(G)| = 1$, then $T_{u,v}(G) = 3$. It is easy to see that $T_{u,v}(G) \geq \omega_{u,v}(G)$ holds for all edges $(u, v) \in E$.

On the computational aspect, computing the edge truss numbers $T_{u,v}(G)$ for all $(u, v) \in E$ has a polynomial time complexity by using an efficeint in-memory algorithm for truss decomposition [54]. Our empirical result shows that in our graph model $\widehat{G} = \widehat{G}_r^{p,q}(\mathcal{X})$, the edge truss number $T_{u,v}$ has the similar



(a) Edge clique number *v.s.* Edge truss number



(b) Edge clique number *v.s.* Degeneracy

Figure 1: In these figures, we compute the edge truss number, the degeneracy [41] (core number) of the (edge) neighborhood graphs, and the edge clique number for all edges in a $(0, 0.01)$-perturbation of the random geometric graph whose vertices are sampled uniformly on a sphere (# nodes: 4957, # edges: 197201). (a) The edge truss number in such graph has the same trend as the edge clique number; (b) The behavior of the degeneracy of the edge neighborhood graph somehow is very different from the edge clique number.

behavior as the edge clique number $\omega_{u,v}$. See Figure 1 for an example. Thus, we conjecture that the edge truss number should be a good candidate for solving the shortest-path metric recovery problem.

Right now I am working on the theoretical perpective of this observed phenomenon.

3.1.2 Filtering by using Ricci curvature

Given a graph $G = (V, E)$, a probability distribution over the vertex set V is a mapping $m : V \to [0, 1]$ such that $\sum_{v \in V} m(v) = 1$. Suppose two probability distributions m_1, m_2 have finite support. A *coupling* between m_1 and m_2 is a mapping $A : V \times V \to [0, 1]$ with the finite support so that

$$\sum_{u \in V} A(u, v) = m_1(v) \text{ and } \sum_{v \in V} A(u, v) = m_2(v)$$

The *graph transportation distance* between m_1 and m_2 is defined as follows.

$$W(m_1, m_2) := \inf_{A} \sum_{u,v \in V} A(u, v) d(u, v)$$

where $d(u, v)$ is the shortest-path distance between u and v in graph G. For any $\alpha \in [0, 1]$ and any vertex $u \in V$, we define the following probability measure m_u^α.

$$m_u^\alpha(v) = \begin{cases} \alpha, & \text{if } v = u, \\ (1 - \alpha)/|N_u(G)|, & \text{if } v \in N_u(G), \\ 0, & \text{otherwise.} \end{cases}$$

Now we are ready to define the α-Ricci curvature of any pair of vertices $u, v \in V$.

Definition 12 (α-Ricci curvature [34])**.** *For any $\alpha \in [0, 1]$ and $u, v \in V$, we define α-Ricci-curvature $\kappa_{u,v}^\alpha$ to be*

$$\kappa_{u,v}^\alpha = 1 - \frac{W\left(m_u^\alpha, m_v^\alpha\right)}{d(u, v)}.$$

As a first step, instead of proving something in noisy random geometric graphs, I am now working on the $\kappa^{\alpha}_{u,v}$ of two different types of edges in stochastic block model [1]. Similar to the (p,q)-perturbation of random geometric graphs, we also randomly remove edges in communities. To be more specific, we are interested in the behavior of $\kappa^{\alpha}_{u,v}$ of edges (u,v) with u,v sitting in the same community (block) as well as those with u,v sitting in different communities. We hope to see a dramatic difference between these two types of edges for reasonable noise level. And if it is true, then a Ricci-curvature based filtering process can be used to differentiate those two types of edges, and it may further be used to design a new method to solve the recovery problem [1].

3.2 Topology

3.2.1 Open problems related to $G_n^{p,q}$

As a new graph model, there are many interesting open problems. For example, the combined case is not yet completely resolved (there are still gaps in the regimes; see Theorem 7 for a concrete example). Also currently we only provide asymptotic tight bounds, and it would be interesting to identify the exact constant for the high order terms too.

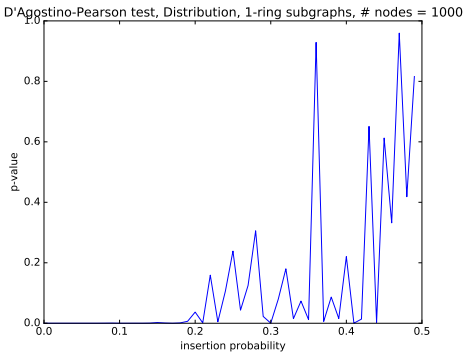
Besides the clique number, the other quantity I would like to study in the future is the *chromatic number* of $G_n^{p,q}$. Given a graph G, the *chromatic number* of G, denoted by $\chi(G)$, is the smallest number of colors needed to color the vertices of G such that each vertex is colored by exact one color, and no two vertices sharing the same edge have the same color. It is well-known that in Erdős–Rényi random graphs $G(n,p)$, $\chi(G(n,p)) = \Theta(n/\log n)$ and $clique(G(n,p)) = \Theta(\log n)$ when p is a constant [3]. Hence, the chromatic number is much larger than the clique number in Erdős–Rényi random graphs $G(n,p)$. However, this is no longer true in random geometric graphs G_n. Interestingly, in random geometric graphs G_n, with high probability, $\chi(G_n)$ the chromatic number is of the same order of $clique(G_n)$ the clique number in all regimes of nr^d [35]. Thus, we are curious about the relationship between the clique number and the chromatic number of $G_n^{p,q}$ (the (p,q)-perturbation of G_n).

3.2.2 Open problems related to $C_{n,p}^{(l)}$

Let $\beta_{1,d}(G) := |\{v \in V : \beta_1(G_{1\to v}) = d\}|$. Inspired by the standard results on the degree distribution of random graphs (Section 3.1 in [17]), a natural question arises: what is the distribution of $\beta_{1,d}(G)$ when G is an Erdős–Rényi random graph $G(n,p)$? Based on our empirical result (see Figure 2), we conjecture that when p is large enough, then $\beta_{1,d}$ should obey a normal distribution. I am now working on proving or disproving this conjecture.

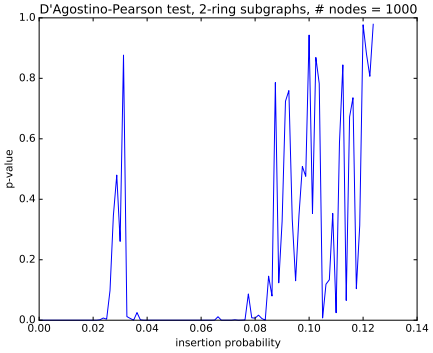
We only considered the 1-ring subgraphs in $G(n,p)$ in this proposal. So how about l-ring subgraphs? Interestingly, for $l = 2, 3$, our empirical result indicates that there should exist two disjoint ranges of p such that when p falls in either range, a central limit theorem of $C_{n,p}^{(l)}$ holds (see Figure 3).

Note that $C_{n,p}^{(l)}$ is the number of (l,l) points in the extended persistence diagrams of rooted subgraphs in Erdős–Rényi random graphs $G(n,p)$. Besides this type of points in the diagrams, we are also interested in other quantities. For example, from the 0-dimensional diagram and 1-dimensional diagram, we can recover the number of vertices, number of edges in each "layer" ($G_{k\to u}$) and the number of crossing edges between two consecutive "layers" (edges between $G_{k\to u}$ and $G_{(k+1)\to u}$). Our next goal is to analyze other quantities in these local diagrams, since these local diagrams may potentially be used to construct a representation of the global graph.

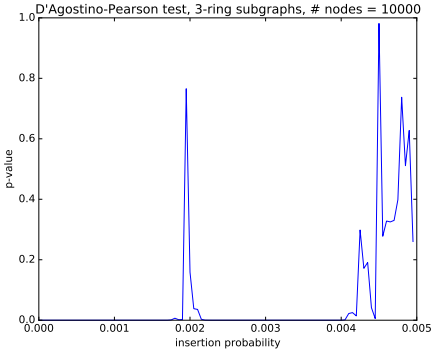


(a) $G(1000, p)$ with $p \in [0, .5]$

Figure 2: Given a graph G sampled from Erdős–Rényi random graph $G(1000, p)$, we perform the so-called D'Agostino-Pearson test [14] on the 1000 samples $\{\beta_1(G_{1\to v}) : v \in G\}$ (one value for each vertex). We simply use the function SCIPY.STATS.NORMALTEST in Python to compute the p-values of these tests (y-axis). Note that a large p-value doesn't mean that the samples are sampled from a normal distribution. However, the result still gives us a hint on how large p should be such that the distribution of $\beta_{1,d}(G)$ looks like a normal distribution.



(a) $G(1000, p)$ with $p \in [0, .125]$



(b) $G(10000, p)$ with $p \in [0, .005]$

Figure 3: Again, we perform the D'Agostino-Pearson tests on 1000 samples of $C^{(2)}_{1000,p}$ and $C^{(3)}_{10000,p}$, respectively. These empirical results don't directly support our conjecture, but they still give us a hint on the range of p in which a central limit theorem of the corresponding 1-dim Betti number holds.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] R. Agrawal, L. de Alfaro, and V. Polychronopoulos. Learning from graph neighborhoods using lstms. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.

[4] P. Bendich, H. Edelsbrunner, D. Morozov, A. Patel, et al. Homology and robustness of level and interlevel sets. *Homology, Homotopy and Applications*, 15(1):51–72, 2013.

[5] H. Bhadauria and M. Dewal. Efficient denoising technique for ct images to enhance brain hemorrhage segmentation. *Journal of digital imaging*, 25(6):782–791, 2012.

[6] O. Bobrowski and M. Kahle. Topology of random geometric complexes: a survey. *Journal of applied and Computational Topology*, 1(3-4):331–364, 2018.

[7] B. Bollobás and P. Erdös. Cliques in random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 80, pages 419–427. Cambridge University Press, 1976.

[8] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press, 2006.

[9] L. Booth, J. Bruck, M. Cook, and M. Franceschetti. Ad hoc wireless networks with noisy links. In *Proceedings of IEEE International Symposium on Information Theory*, pages 386–386. IEEE, 2003.

[10] M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2011.

[11] D. Coppersmith, D. Gamarnik, and M. Sviridenko. *The Diameter of a Long-Range Percolation Graph*, pages 147–159. Birkhäuser Basel, Basel, 2002.

[12] F. Costa and K. De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262. Omnipress; Madison, WI, USA, 2010.

[13] N. R. Council et al. *Mathematical challenges from theoretical/computational chemistry*. National Academies Press, 1995.

[14] R. D'AGOSTINO and E. S. Pearson. Tests for departure from normality. *Biometrika*, 60(3):613–622, 1973.

[15] T. K. Dey, D. Shi, and Y. Wang. Comparing graphs via persistence distortion. *arXiv preprint arXiv:1503.07414*, 2015.

[16] R. Durrett. *Random Graph Dynamics*, volume 20. Cambridge University Press, 2006.

[17] A. Frieze and M. Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[18] D. S. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, 100(8):4372–4376, 2003.

[19] G. R. Grimmett and C. J. McDiarmid. On colouring random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 77, pages 313–324. Cambridge University Press, 1975.

[20] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005.

[21] J. Heinonen. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2012.

[22] I. Iacopini, G. Petri, A. Barrat, and V. Latora. Simplicial models of social contagion. *Nature communications*, 10(1):2485, 2019.

[23] S. Janson, R. Kozma, M. Ruszinkó, and Y. Sokolov. Bootstrap percolation on a random graph coupled with a lattice. *ELECTRONIC JOURNAL OF COMBINATORICS*, 2016.

[24] M. Kahle. Sharp vanishing thresholds for cohomology of random flag complexes. *Annals of Mathematics*, pages 1085–1107, 2014.

[25] M. Kahle, M. Tian, and Y. Wang. Local cliques in er-perturbed random geometric graphs. *arXiv preprint arXiv:1810.08383*, 2018. To be appeared at ISAAC 2019.

[26] M. Kahle, M. Tian, and Y. Wang. On the clique number of noisy random geometric graphs. Submitted, 2019.

[27] A. P. Kartun-Giles and G. Bianconi. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *Chaos, Solitons & Fractals: X*, 1:100004, 2019.

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

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

[30] J. Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1019–1044, 2006.

[31] N. M. Kriege, F. D. Johansson, and C. Morris. A survey on graph kernels. *arXiv preprint arXiv:1903.11835*, 2019.

[32] T. Lei, W. Jin, R. Barzilay, and T. Jaakkola. Deriving neural architectures from sequence and graph kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2024–2033. JMLR. org, 2017.

[33] E. A. Leicht, P. Holme, and M. E. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[34] Y. Lin, L. Lu, and S.-T. Yau. Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627, 2011.

[35] C. McDiarmid and T. Müller. On the chromatic number of random geometric graphs. *Combinatorica*, 31(4):423–488, 2011.

[36] R. Meester and R. Roy. *Continuum percolation*, volume 119. Cambridge University Press, 1996.

[37] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[38] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[39] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019.

[40] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu, and S. Saminathan. subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs. *arXiv preprint arXiv:1606.08928*, 2016.

[41] M. Newman. *Networks*. Oxford university press, 2018.

[42] M. Niepert, M. Ahmed, and K. Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.

[43] G. Nikolentzos, G. Siglidis, and M. Vazirgiannis. Graph kernels: A survey. *arXiv preprint arXiv:1904.12218*, 2019.

[44] S. Oudot and E. Solomon. Barcode embeddings for metric graphs. *arXiv preprint arXiv:1712.03630*, 2017.

[45] S. Parthasarathy, D. Sivakoff, M. Tian, and Y. Wang. A quest to unravel the metric structure behind perturbed networks. In *33rd International Symposium on Computational Geometry (SoCG 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[46] M. Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

[47] G. W. Peters and T. Matsui. *Theoretical Aspects of Spatial-Temporal Modeling*. Springer, 2015.

[48] V. Satuluri, S. Parthasarathy, and Y. Ruan. Local graph sparsification for scalable clustering. In *ACM SIGMOD Intl. Conf. Management Data*, pages 721–732, 2011.

[49] 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.

[50] 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.

[51] S. Srihari, C. H. Yong, and L. Wong. *Computational Prediction of Protein Complexes from Protein Interaction Networks*. Morgan & Claypool, 2017.

[52] D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter, and P. J. Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature communications*, 6:7723, 2015.

[53] M. Tian and Y. Wang. A limit theorem for the 1st betti number of 1-ring subgraphs in random graphs. In preparation.

[54] J. Wang and J. Cheng. Truss decomposition in massive networks. *Proceedings of the VLDB Endowment*, 5(9):812–823, 2012.

[55] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296(5571):1302–1305, 2002.

[56] D. J. Watts and S. H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[57] X. Y. Wu. Mixing time of random walk on poisson geometry small world. *Internet Mathematics*, 2017.

[58] C. Yang, M. Liu, V. W. Zheng, and J. Han. Node, motif and subgraph: Leveraging network functional blocks through structural convolution. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 47–52. IEEE, 2018.

[59] M. Zhang and Y. Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175, 2018.

[60] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.