Local Cliques in ER-Perturbed Random Geometric Graphs

Matthew Kahle

Department of Mathematics, The Ohio State University, USA
mkahle@math.osu.edu

Minghao Tian

Computer Science and Engineering Dept., The Ohio State University, USA
tian.394@osu.edu

Yusu Wang

Computer Science and Engineering Dept., The Ohio State University, USA
yusu@cse.ohio-state.edu

Abstract

We study a random graph model introduced in [20] where one adds Erdős–Rényi (ER) type perturbation to a random geometric graph. More precisely, assume $G^*_{\mathcal{X}}$ is a random geometric graph sampled from a nice measure on a metric space $\mathcal{X} = (X, d)$. An *ER-perturbed random geometric graph* $\widehat{G}(p, q)$ is generated by removing each existing edge from $G^*_{\mathcal{X}}$ with probability p, while inserting each non-existent edge to $G^*_{\mathcal{X}}$ with probability q. We consider a localized version of clique number for $\widehat{G}(p, q)$: Specifically, we study the *edge clique number* for each edge in a graph, defined as the size of the largest clique(s) in the graph containing that edge. We show that the edge clique number presents two fundamentally different types of behaviors in $\widehat{G}(p, q)$, depending on which "type" of randomness it is generated from.

As an application of the above results, we show that by a simple filtering process based on the edge clique number, we can recover the shortest-path metric of the random geometric graph $G^*_{\mathcal{X}}$ within a multiplicative factor of 3 from an ER-perturbed observed graph $\widehat{G}(p, q)$, for a significantly wider range of insertion probability q than what is required in [20].

2012 ACM Subject Classification Mathematics of computing → Random graphs; Theory of computation → Computational geometry

Keywords and phrases random graphs, random geometric graphs, edge clique number, the probabilistic method, metric recovery

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2019.29

Related Version A full version of the paper is available at `https://arxiv.org/abs/1810.08383`.

1 Introduction

Random graphs are mathematical models which have applications in a wide spectrum of domains. Erdős–Rényi graph $G(n, p)$ is one of the oldest and most-studied models for networks [19], constructed by adding edges between all pairs of n vertices with probability p independently. Many global properties of this model are well-studied by using the probabilistic method [1], such as the clique number and the phase transition behaviors of connected components w.r.t. parameter p.

Another classical type of random graphs is the random geometric graph $G(\mathbb{X}_n; r)$ introduced by Edgar Gilbert in 1961 [10]. This model starts with a set of n points $\mathbb{X}_n$ randomly sampled over a metric space (typically a cube in $\mathbb{R}^d$) from some probability distribution, and edges are added between all pairs of points within distance r to each other. The Erdős–Rényi random graphs and random geometric graphs exhibit similar behavior for the Poisson degree distribution; however, other properties, such as the clique number and phase transition (w.r.to p or to r), could be very different [11, 16, 21, 22].

This model has many applications in real world where the physical locations of objects involved play an important role [7], for example wireless ad hoc networks [18] and transportation networks [2].

We are interested in mixed models that "combine" both types of randomness together. One way to achieve this is to add Erdős–Rényi type perturbation (percolation) to random geometric graphs. A natural question arises: what are the properties of this type of random graphs? Although these graphs are related to the continuum percolation theory [17], our understanding about them so far is still limited: In previous studies, the underlying spaces are typically plane (called the Gilbert disc model) [3], cubes [6] and tori [13]; the vertices are often chosen as the standard lattices of the space; and the results usually concern the connectivity [4, 23] or diameter [24].

Our work. In this paper, we consider a mixed model of Erdős–Rényi random graphs and random geometric graphs, and study the behavior of a local property called *edge clique number*. More precisely, we use the following *ER-perturbed random geometric graph model* previously introduced in [20]. Suppose there is a compact metric space $\mathcal{X} = (X, d)$ (as feature space) with a probability distribution induced by a "nice" measure μ supported on $\mathcal{X}$ (e.g., the uniform measure supported on an embedded smooth low-dimensional Riemannian manifold). Assume we now randomly sample n points V i.i.d from this measure μ, and build the random geometric graph $G^*_{\mathcal{X}}(r)$, which is the r-neighborhood graph spanned by V (i.e, two points $u, v \in V$ are connected if their distance $d(u, v) \le r$). Next, we add Erdős–Rényi (ER) type perturbation to $G^*_{\mathcal{X}}(r)$: each edge in $G^*_{\mathcal{X}}(r)$ is deleted with a uniform probability p, while each "short-cut" edge between two unconnected nodes u, v is inserted to $G^*_{\mathcal{X}}(r)$ with a uniform probability q. We denote the resulting generated graph by $\widehat{G}^{p;q}_{\mathcal{X}}(r)$.

Intuitively, one can imagine that a graph is generated first as a proximity graph (captured by the random geometric graph) in some feature space ($\mathcal{X}$ in the above setting). The random insertion / deletion of edges then allows for noise or exceptions. For example, in a social network, nodes could be sampled from some feature space of people, and two people could be connected if they are nearby in the feature space. However, there are always some exceptions – friends could be established by chance even they are very different from each other ("far away"), and two similar ("close") people (say, close geographically and in tastes) may not develop friendship. The ER-perturbation introduced above by [20] aims to account for such kind of exceptions.

We introduce a local property called the *edge clique number* of a graph G, to provide a more refined view than the global clique number. It is defined for each edge (u, v) in the graph, denoted as $\omega_{u,v}(G)$, as the size of the largest clique containing uv in graph G. Our main result is that $\omega_{u,v}\left(\widehat{G}^{p;q}_{\mathcal{X}}(r)\right)$ presents two fundamentally different types of behaviors, depending on from which "type" of randomness the edge (u, v) is generated from: A "good" edge from the random geometric graph $G^*_{\mathcal{X}}(r)$ has an edge-clique number similar to edges from a certain random geometric graph; while a "bad" edge (u, v) introduced during the random-insertion process has an edge-clique number similar to edges in some random Erdős–Rényi graph. See Theorems 5, 10, 12, and 14 for the precise statements.

As an application of our theoretical analysis, in Theorem 15, we show that by using a filtering process based on our edge clique number, we can recover the shortest-path metric of the random geometric graph $G^*_{\mathcal{X}}(r)$ within a multiplicative factor of 3, from an ER-perturbed graph $\widehat{G}^{p;q}_{\mathcal{X}}(r)$, for a significantly wider range of insertion probability q than what's required in [20], although we do need a stronger regularity condition on the measure μ. See more discussion at the end of Section 4.

2 Preliminaries

Suppose we are given a compact geodesic metric space $\mathcal{X} = (X, d)$ [5] [1]. We will consider "nice" measures on $\mathcal{X}$. Specifically,

▶ **Definition 1** (Doubling measure [12])**.** *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 $\mu : X \to \mathbb{R}$ 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. For our theoretical results later, we in fact need a stronger condition on the input measure, which we will specify later in Assumption-A at the beginning of Section 3.

ER-perturbed random geometric graph. Following [20], 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 $G^*_{\mathcal{X}}(r) = (V, E^*)$ for some parameter $r > 0$ on V; that is, $E^* = \{(u, v) \mid d(u, v) \leq r, u, v \in V\}$. We call $G^*_{\mathcal{X}}(r)$ a random geometric graph 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 $\widehat{G}^{p,q}_{\mathcal{X}}(r) = (V, \widehat{E})$ is called a *(p, q)-perturbation of $G^*_{\mathcal{X}}(r)$*, or simply an *ER-perturbed random geometric graph.*

We now introduce a local version of the standard clique number:

▶ **Definition 2** (Edge clique number)**.** *Given a graph $G = (V, E)$, for any edge $(u, v) \in E$, its* edge clique number *$\omega_{u,v}(G)$ is defined as*

$$\omega_{u,v}(G) = \text{ the size of the largest clique(s) in } G \text{ containing } (u, v).$$

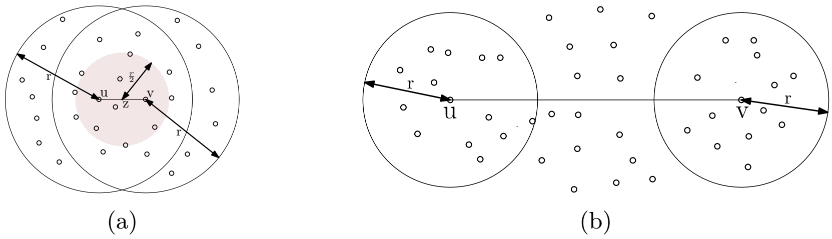
Setup for the remainder of the paper. For convenience of reference, we collect our standard notations. We assume throughout that we are given a fixed compact geodesic metric space $\mathcal{X} = (X, d)$ and a fixed L-doubling probability measure μ. We denote V as the set of n graph nodes sampled i.i.d. from μ. $\widehat{G} = \widehat{G}^{p,q}_{\mathcal{X}}(r) = (V, \widehat{E})$ is a (p, q)-perturbation of a random geometric graph $G^* = G^*_{\mathcal{X}}(r)$ spanned by V with radius parameter r. For an arbitrary graph G, let $V(G)$ and $E(G)$ refer to its vertex set and edge set, respectively, and 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 $E(G)$).

We now define two types of edges in the perturbed graph $\widehat{G}$. Roughly speaking, we say an edge in $\widehat{G}$ is a good-edge if it is generated by the random geometric graph $G*$ and later is not removed by $(p, q)-$perturbation. A bad-edge is typically some long-range edge inserted by the perturbation.

[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 3** (Good / bad-edges). *An edge (u, v) in the perturbed graph $\widehat{G}$ is a* good-edge *if $d(u, v) \leq r$. An edge (u, v) in the perturbed graph $\widehat{G}$ is a* bad-edge *if for any $x \in N_{G^*}(u)$ and $y \in N_{G^*}(v)$, we have $d(x, y) > r$.*

In other word, (u, v) is a bad-edge if and only if there are no edges between neighbors of u and neighbors of v in G^*. See figure 1 for some examples.



■ **Figure 1** (a) shows a good-edge (u, v). It also shows that if $d(u, v) \leq r$, then there exists an $r/2$-ball (shaded region) in the intersection of $B_r(u)$ and $B_r(v)$; (b) shows a bad-edge (u, v).

Organization of paper. In Section 3, we study the behavior of edge clique number for good-edges and bad-edges. Our main result Theorem 14 roughly suggests, under certain conditions on the insertion probability q, for a good-edge (u, v) of $\widehat{G}^{p,q}_{\mathcal{X}}(r)$, with high probability, $\omega_{u,v}\left(\widehat{G}\right)$ has order $\Omega\left(\ln\ln n\right)$; while for a bad-edge (u, v), its edge-clique number $\omega_{u,v}\left(\widehat{G}\right)$ has order $o\left(\ln\ln n\right)$ with high probability.

To illustrate the main ideas, we will first give results for when only edge-insertion type of perturbations is added to the random geometric graph in Section 3.1 – In fact, this case is of independent interest as well. An application of our result to recover the shortest-path metric of the hidden geometric graph is given in Section 4.

3 Two different behaviors of edge clique number

In Section 3.1, we study the edge clique numbers for the insertion-only perturbed random geometric graphs, both to illustrate the main ideas, and to show the different behaviors of the edge clique number more clearly. In Section 3.2, we study the case for deletion-only perturbed random geometric graphs, where we only delete each edge independently with probability p to obtain an input graph $\widehat{G}$. Finally, we discuss the combined case of an ER-perturbed random geometric graph in Section 3.3.

First, we need the following technical assumption on the parameter r (for the random geometric graph $G^*_{\mathcal{X}}(r)$) and the measure μ where graph nodes V are sampled from.

[Assumption-A]: 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) $\mu\left(B_{r/2}(x)\right) \leq \rho\mathrm{s}$

Intuitively, these two conditions require that for the specific r value we choose, the mass contained inside all radius-r metric balls are similar (within a constant ρ factor). Density-cond is equivalent to the Assumption-R in [20]. It requires that r is large enough such that with high probability each vertex v in the random geometric graph $G^*_{\mathcal{X}}(r)$ has degree $\Omega(\ln n)$. Indeed, we have the following claim.

▷ Claim 4 ([20]). Under Density-cond, with probability at least $1 - n^{-5/3}$, each vertex in $G^*_{\mathcal{X}}(r)$ has at least $sn/4$ neighbors.

Proof. For a fixed vertex $v \in V$, let n_v be the number of points in $(V - \{v\}) \cap \mathrm{B}_r(v)$. The expectation of n_v is $(n-1) \cdot \mu\left(\mathrm{B}_r(v)\right) \geq (n-1) \cdot \mu\left(B_{\frac{r}{2}}(v)\right) \geq s(n-1)$. By the Chernoff bound, we thus have that

$$\mathbb{P}\left[n_v < \frac{sn}{4}\right] < \mathbb{P}\left[n_v < \frac{s(n-1)}{3}\right] \leq \mathbb{P}\left[n_v < \frac{1}{3}(n-1)\mu\left(\mathrm{B}_r(v)\right)\right]$$
$$\leq e^{-\frac{\left(\frac{2}{3}\right)^2}{2}(n-1)\mu(\mathrm{B}_r(v))} \leq n^{-\frac{8}{3}}$$

It then follows from the union bound that the probability that all n vertices in V have more than $sn/4$ neighbors is at least $1 - n \cdot n^{-\frac{8}{3}} = 1 - n^{-5/3}$. ◁

3.1 Insertion-only perturbation

Recall $G^*_{\mathcal{X}}(r) = (V, E)$ is a random geometric graph whose n vertices V sampled i.i.d. from a L-doubling probability measure μ supported on a compact metric space $\mathcal{X} = (X, d)$. In this section, we assume that the input graph $\widehat{G}$ is generated from $G^* = G^*_{\mathcal{X}}(r)$ as follows: First, include all edges of G^* in $\widehat{G}$. Next, for any $u, v \in V$ with $(u, v) \neq E(G^*)$, we add edge (u, v) to $E(\widehat{G})$ with probability q. That is, we only insert edges to G^* to obtain $\widehat{G}$.

First, for good-edges, it is easy to obtain the following result.

▶ **Theorem 5.** *Assume Density-cond holds. Let G^* be an n-node random geometric graph generated from (X, d, μ) as described. Denote $\widehat{G} = \widehat{G}^q$ the final graph after inserting each edge not in G^* independently with probability q. Then, with high probability, for each good-edge (u, v) in $\widehat{G}$, its edge clique number satisfies that $\omega_{u,v}(\widehat{G}) \geq \mathrm{s}n/4$.*

Proof. For each good-edge (u, v), observe that $B_r(u) \cap B_r(v)$ contains at least one metric ball of radius $r/2$ (say $B_{r/2}(z)$ with z being the mid-point of a geodesic connecting u to v in X, see Figure 1 (a)). And all the points in an $r/2-$ball span a clique in G^* ($r-$neighborhood graph). Then by an argument similar to the proof of Claim 4, we have that with probability at least $1 - n^{-\frac{2}{3}}$, the number of points in all of $O(n^2)$ number $r/2$-balls centered at some mid-point of the geodesics between all pair of nodes $u, v \in V$ is at least $\mathrm{s}n/4$. Hence with probability at least $1 - n^{-\frac{2}{3}}$, for all good-edge (u, v) in $\widehat{G}$, $\omega_{u,v}(\widehat{G}) \geq \mathrm{s}n/4$. ◀

Bounding the edge clique number for bad-edges is much more challenging due to the interaction between local edges (from random geometric graph) and long-range edges (from random insertion). To handle this, we will create a finite specific collection of subgraphs for $\widehat{G}$ in an appropriate manner, and bound the edge clique number of a bad-edge in each such subgraph. The property of this specific collection of subgraphs is that the union of these individual cliques provides an upper bound on the edge clique number for this edge in $\widehat{G}$. To construct this finite collection of subgraphs, we will use the so-called Besicovitch covering lemma which has a lot of applications in measure theory [8]. The finiteness here is crucial for later applying the union bound (i.e., Bonferroni inequality [9]).

First, we introduce some notations. We use a *packing* to refer to a countable collection $\mathcal{B}$ of *pairwise disjoint* closed balls. Such a collection $\mathcal{B}$ is a *packing w.r.t. a set P* if the centers of the balls in $\mathcal{B}$ lie in the set $P \subset X$, and it is a *δ-packing* if all of the balls in $\mathcal{B}$ have radius δ. A set $\{A_1, \ldots, A_\ell\}$, $A_i \subseteq X$, *covers P* if $P \subseteq \bigcup_i A_i$.

▶ **Lemma 6** (Besicovitch Covering Lemma, doubling space version, [14]). *Let $\mathcal{X} = (X, d)$ be a doubling space. Then, there exists a constant $\beta = \beta(\mathcal{X}) \in \mathbb{N}$ such that for any $P \subset X$ and $\delta > 0$, there are β different δ-packings w.r.t. P, denoted by $\{\mathcal{B}_1, \cdots, \mathcal{B}_\beta\}$, whose union also covers P.*

We call the constant $\beta(\mathcal{X})$ above the *Besicovitch constant*. Note that this constant only depends on the doubling space $\mathcal{X}$ and thus is finite. Given a set A, we say that A is *partitioned into* $A_1, A_2, \cdots, A_k$, if $A = A_1 \cup \cdots \cup A_k$ and $A_i \cap A_j = \emptyset$ for any $i \neq j$.

▶ **Definition 7** (Well-separated clique-partitions family). *Consider the random geometric graph $G^* = G^*_{\mathcal{X}}(r)$. A family $\mathcal{P} = \{P_i\}_{i\in\Lambda}$, where $P_i \subseteq V$ and Λ is the index set of P_is, forms a* well-separated clique-partitions family *of G^* if:*

1. $V = \cup_{i\in\Lambda} P_i$.
2. $\forall i \in \Lambda$, *P_i can be partitioned as $P_i = C_1^{(i)} \sqcup C_2^{(i)} \sqcup \cdots \sqcup C_{m_i}^{(i)}$ where*

(2-a) $\forall j \in [1, m_i]$, *there exist $\bar{v}_j^{(i)} \in V$ such that $C_j^{(i)} \subseteq B_{r/2}\left(\bar{v}_j^{(i)}\right) \cap V$.*

(2-b) For any $j_1, j_2 \in [1, m_i]$ with $j_1 \neq j_2$, $d_H\left(C_{j_1}^{(i)}, C_{j_2}^{(i)}\right) > r$, where d_H is the Hausdorff distance between two sets in metric space (X, d).

We also call $C_1^{(i)} \sqcup C_2^{(i)} \sqcup \cdots \sqcup C_{m_i}^{(i)}$ a clique-partition of P_i (w.r.t. G^*)*, and its* size *(cardinality) is m_i. The* size *of the well-separated clique-partitions family $\mathcal{P}$ is its cardinality $|\mathcal{P}| = |\Lambda|$.*

In the above definition, (2-a) implies that each $C_j^{(i)}$ spans a clique in G^*; thus we call $C_j^{(i)}$ a *clique* in P_i and $C_1^{(i)} \sqcup C_2^{(i)} \sqcup \cdots \sqcup C_{m_i}^{(i)}$ a clique-partition of P_i. (2-b) means that there are no edges in G^* between any two cliques of P_i; thus, any edge in $\widehat{G}$ between such cliques must come from insertion. The following existence lemma can be derived by applying Lemma 6 several times.

▶ **Lemma 8.** *There is a well-separated clique-partitions family $\mathcal{P} = \{P_i\}_{i\in\Lambda}$ of $G^*_{\mathcal{X}}(r)$ with $|\Lambda| \leq \beta^2$, where $\beta = \beta(\mathcal{X})$ is the Besicovitch constant of $\mathcal{X}$.*

Proof. To prove the lemma, first we grow an $r/2$-ball around each node in $V \subset X$ (the vertex set of G^*). By Besicovitch covering lemma (Lemma 6), we have a family of $(r/2)$-packings w.r.t. V, $\mathcal{B} = \{\mathcal{B}_1, \cdots, \mathcal{B}_{\alpha_1}\}$, whose union covers V. Here, the constant α_1 satisfies $\alpha_1 \leq \beta(\mathcal{X})$.

Each $\mathcal{B}_i$ contains a collection of disjoint $r/2$-balls centered at a subset of nodes in V, and let $V_i \subseteq V$ denote the centers of these balls. For any $u, v \in V_i$, we have $d(u, v) > r$ as otherwise, $B_{r/2}(u) \cap B_{r/2}(v) \neq \emptyset$ meaning that the $r/2$-balls in $\mathcal{B}_i$ are not all pairwise disjoint. Now consider the collection of r-balls centered at all nodes in V_i. Applying Besicovitch covering lemma to V_i again with $\delta = r$, we now obtain a family of r-packings w.r.t. V_i, denoted by $\mathcal{D}^{(i)} = \mathcal{D}_1^{(i)} \sqcup \cdots \sqcup \mathcal{D}_{\alpha_2^{(i)}}^{(i)}$, whose union covers V_i. Here, the constant $\alpha_2^{(i)}$ satisfies $\alpha_2^{(i)} \leq \beta(\mathcal{X})$ for each $i \in [1, \alpha_1]$.

Now each $\mathcal{D}_j^{(i)}$ contains a set of disjoint r-balls centered at a subset of nodes $V_j^{(i)} \subseteq V_i$ of V_i. First, we claim that $\bigcup_j V_j^{(i)} = V_i$. This is because that $\mathcal{B}_i$ is an $r/2$-packing which implies that $d(u, v) > r$ for any two nodes $u, v \in V_i$. In other words, the r-ball around any node from V_i contains no other nodes in V_i. As the union of r-balls $\mathcal{D}_1^{(i)} \sqcup \cdots \sqcup \mathcal{D}_{c_2^{(i)}}^{(i)}$ covers V_i by construction, it is then necessary that each node V_i has to appear as the center in at least one $\mathcal{D}_j^{(i)}$ (i.e, in $V_j^{(i)}$). Hence $\bigcup_j V_j^{(i)} = V_i$.

Now for each vertex set $V_j^{(i)}$, let $P_j^{(i)} \subseteq V$ denote all points from V contained in the $r/2$-balls centered at points in $V_j^{(i)}$. As $\cup_j V_j^{(i)} = V_i$, we have that $\bigcup_j P_j^{(i)} = \bigcup_{v \in V_i} \left(B_{r/2}(v) \cap V\right)$. It then follows that $\bigcup_{i\in[1,\alpha_1]} \left(\bigcup_{j\in[1,\alpha_2^{(i)}]} P_j^{(i)}\right) = V$ as the union of the family of $r/2$-packings $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_{c_1}\}$ covers all points in V (recall that $\mathcal{B}_i$ is just the set of $r/2$-balls centered at points in V_i).

Clearly, each $P_j^{(i)}$ adapts a clique-partition: Indeed, for each $V_j^{(i)}$, any two nodes in $V_j^{(i)}$ are at least distance $2r$ apart (as the r-balls centered at nodes in V_j^i are disjoint), meaning that the $r/2$-balls around them are more than r (Hausdorff-)distance away. In other words, $\mathcal{P} = \left\{P_j^{(i)}, i \in [1,\alpha_1], j \in [1,\alpha_2^{(i)}]\right\}$ forms a well-separated clique-partitions family of G^*. Finally, since $\alpha_1, \alpha_2^{(i)} \le \beta(\mathcal{X}) = \beta$, the cardinality of $\mathcal{P}$ is thus bounded by β^2. ◀

We also need the following lemma to upper-bound the number of points in every $r/2$-ball centered at nodes of G^*.

▶ **Lemma 9.** *Suppose $G^* = (V, E^*)$ is an n-node random geometric graph sampled from $(\mathcal{X}, \mu, r)$. If Assumption-A holds, then with probability at least $1 - n^{-5}$, for every $v \in V$, the ball $B_{r/2}(v) \cap V$ contains at most $3\rho sn$ points.*

Proof. For a fixed vertex $v \in V$, let $n_{v,r/2}$ be the number of points in $(V - \{v\}) \cap B_{r/2}(v)$. By the definition of random geometric graph, we know that $n_{v,r/2}$ is subject to binomial distribution $Bin\left(n-1, \mu\left(B_{r/2}(v)\right)\right)$. The expectation of $n_{v,r/2}$ is $(n-1)\mu(B_{r/2}(v)) \le \rho sn$. Also note that $(n-1)\mu\left(B_{r/2}(v)\right) \ge (n-1)s \ge 12\ln n$. By applying the Chernoff bound, we thus have that

$$\mathbb{P}\left[n_{v,r/2} \ge \frac{5}{2}\rho sn\right] \le \mathbb{P}\left[n_{v,r/2} \ge \frac{5}{2}(n-1)\mu\left(B_{r/2}(v)\right)\right] \le e^{-\frac{1}{3}\left(\frac{3}{2}\right)(n-1)\mu\left(B_{r/2}(v)\right)} \le n^{-6}$$

Finally, by applying the union bound, we know that with probability at least $1 - n \cdot n^{-6} = 1 - n^{-5}$, $\forall v \in G^*$, there are at most $\frac{5}{2}\rho sn + 1 < 3\rho sn$ points in the geodesic ball $B_{r/2}(v)$. ◀

We now state one of our main theorems, which relates the edge clique number for bad-edges with the insertion probability. To simplify notations, we call a clique containing an edge (u, v) a *uv-clique*.

▶ **Theorem 10.** *Assume Assumption-A holds. Let $\widehat{G} = \widehat{G}^q$ denote the graph obtained by inserting each edge not in $G^*_{\mathcal{X}}(r)$ independently with probability q. Then there exist constants $c_1, c_2, c_3 > 0$ which depend on the doubling constant L of μ, the Besicovitch constant $\beta(\mathcal{X})$, and the regularity constant ρ, such that for any $\mathsf{K} = \mathsf{K}(n)$ with $\mathsf{K} \to \infty$ as $n \to \infty$, with high probability, $\omega_{u,v}(\widehat{G}) < \mathsf{K}$ for any bad-edge (u, v) in $\widehat{G}$, as long as q satisfies*

$$q \quad \le \quad \min\left\{ c_1, \quad c_2 \cdot \left(\frac{1}{n}\right)^{c_3/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn} \right\}. \tag{1}$$

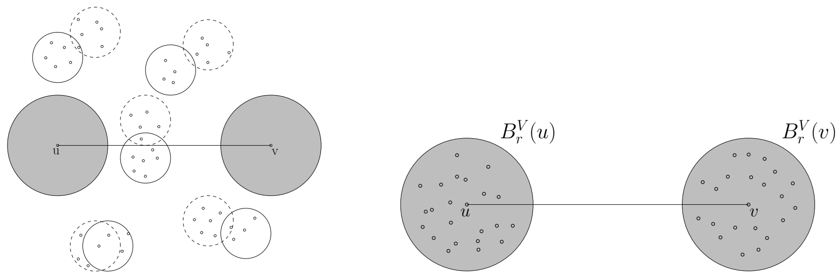
▶ Remark. To illustrate the above theorem, consider for example when $\mathsf{K} = \Theta(sn)$. Then the theorem says that there exists constant c' such that if $q < c'$, then w.h.p. $\omega_{u,v} < \mathsf{K}$ (thus $\omega_{u,v} = O(sn)$) for any bad-edge (u, v). Now consider when $q = o(1)$. Then the theorem implies that w.h.p. the edge-clique number for any bad-edge is at most $\mathsf{K} = o(sn)$. This is qualitatively different from the edge-clique number for a good-edge for the case $q = o(1)$, which is $\Omega(sn)$ as shown in Theorem 5. By reducing this insertion probability q, this gap can be made **larger and larger**.

Proof of Theorem 10. Given any node y, let $B_r^V(y) \subseteq V$ denote $B_r(y) \cap V$. Now consider a bad-edge (u, v). Set $A_{uv} = \{w \in V | w \notin B_r(u) \cup B_r(v)\}$ and $B_{uv} = \{w \in V | w \in B_r^V(u) \cup B_r^V(v)\}$. Denote $\tilde{A}_{uv} = A_{uv} \cup \{u\} \cup \{v\}$; It is easy to check that $V = \tilde{A}_{uv} \cup B_{uv}$.

Let $G|_S$ denote the subgraph of G spanned by a subset S of its vertices. Given any set C, let $\mathrm{C}|_S = \mathrm{C} \cap S$ be the restriction of C to another set S. Now consider a subset of vertices $\mathrm{C} \subseteq V$: obviously, $\mathrm{C} = \mathrm{C}|_{\tilde{A}_{uv}} \cup \mathrm{C}|_{B_{uv}}$. Hence by the pigeonhole principle and the union bound, we have:

$$\begin{aligned}
&\mathbb{P}\left[\widehat{G} \text{ has a } uv\text{-clique of size} \geq \mathsf{K}\right] \\
&\leq \mathbb{P}\left[\widehat{G}|_{\tilde{A}_{uv}} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2}\right] + \mathbb{P}\left[\widehat{G}|_{B_{uv}} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2}\right] \qquad (2)
\end{aligned}$$

Next, we will bound the two terms on the right hand side of Eqn. (2) separately in Case (A) and Case (B) below.



(a) (b)

Figure 2 (a) A well-separated clique partition $\mathcal{P} = \{P_1, P_2\}$ of A_{uv} – points in the solid ball are P_1, and those in dashed ball are P_2. (b) Points in B_{uv}.

Case (A): bounding the first term in Eqn. (2). We apply Lemma 8 for points in A_{uv}. This gives us a well-separated clique-partitions family $\mathcal{P} = \{P_i\}_{i \in \Lambda}$ of A_{uv} with $|\Lambda|$ being a constant (see Figure 2 (a)). Augment each P_i to $\tilde{P}_i = P_i \cup \{u\} \cup \{v\}$. Suppose there is a clique C in $\widehat{G}|_{\tilde{A}_{uv}}$, then as $\bigcup_i \tilde{P}_i = \tilde{A}_{uv}$, we have $C = \bigcup_{i \in \Lambda} C|_{\tilde{P}_i}$, implying that $|C| \leq \sum_{i \in \Lambda} \left|C|_{\tilde{P}_i}\right|$. Hence by pigeonhole principle and the union bound, we have:

$$\mathbb{P}\left[\widehat{G}|_{\tilde{A}_{uv}} \text{has a} uv\text{-clique of size} \geq \frac{\mathsf{K}}{2}\right] \leq \sum_{i=1}^{|\Lambda|} \mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{has a} uv\text{-clique of size} \geq \frac{\mathsf{K}}{2|\Lambda|}\right] \qquad (3)$$

Now for arbitrary $i \in \Lambda$, consider $\widehat{G}|_{\tilde{P}_i}$, the induced subgraph of $\widehat{G}$ spanned by vertices in $\tilde{P}_i$. Note, $\widehat{G}|_{\tilde{P}_i}$ can be viewed as generated by inserting each edge not in $G^*|_{\tilde{P}_i} \cup \{uv\}$ to it with probability q. Recall from Definition 7 that each P_i adapts a clique decomposition $C_1^{(i)} \sqcup \cdots \sqcup C_{m_i}^{(i)}$, where every $C_j^{(i)}$ is contained in an $r/2$-ball, and all such balls are r-separated (w.r.t Hausdorff distance).

Fix any $i \in \Lambda$. For simplicity of the argument below, set $m = m_i$, and let $N_j = \left|C_j^{(i)}\right|$ denote the number of points in the j-th cluster $C_j^{(i)}$. Note that obviously, $m \leq |P_i| \leq |V| = n$ for any $i \in \Lambda$. Set $N_{\max} = 3\rho s n$. By Lemma 9, we know that, with high probability (at least $1 - n^{-5}$), $N_j \leq N_{\max}$ for all j in $[1, m]$. Let F denote the event that "for every $v \in V$, the ball $B_{r/2}(v) \cap V$ contains at most $N_{\max}$ points"; and F^c denotes the complement event of F.

Now set $k := \left\lfloor \frac{\mathsf{K}}{2|\Lambda|} \right\rfloor - 2$. For every set S of $k+2$ vertices in this graph $\widehat{G}|_{\tilde{P}_i}$, let A_S be the event "*S is a uv-clique in $\widehat{G}|_{\tilde{P}_i}$*" and X_S its indicator random variable. Set $\mathrm{X} = \sum_{|S|=k+2} \mathrm{X}_S$ and note that X is the number of uv-cliques of size $(k+2)$ in $\widehat{G}|_{\tilde{P}_i}$. It follows from Markov inequality that:

$$\begin{aligned}\mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{ has a } uv\text{-clique of size} \geq k+2\right] &= \mathbb{P}[\mathrm{X} > 0] \leq \mathbb{P}[\mathrm{X} > 0 \mid \mathrm{F}] + \mathbb{P}[\mathrm{F}^c] \\ &\leq \mathbb{E}[\mathrm{X} \mid \mathrm{F}] + n^{-5}. \end{aligned} \tag{4}$$

On the other hand, using linearity of expectation, we have:

$$\begin{aligned}\mathbb{E}[\mathrm{X} \mid \mathrm{F}] = \sum_{|S|=k+2} \mathbb{E}[\mathrm{X}_S \mid \mathrm{F}] &= q^{2k} \sum_{\substack{x_1+x_2+\cdots+x_m=k \\ 0\leq x_i \leq N_i}} \binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_m}{x_m} q^{(k^2-\sum_{i=1}^m x_i^2)/2} \\ &\leq q^{2k} \sum_{\substack{x_1+x_2+\cdots+x_m=k \\ 0\leq x_i \leq N_{\max}}} \binom{N_{\max}}{x_1}\binom{N_{\max}}{x_2}\cdots\binom{N_{\max}}{x_m} q^{(k^2-\sum_{i=1}^m x_i^2)/2} \end{aligned} \tag{5}$$

To estimate this quantity, we have the following lemma:

▶ **Lemma 11.** *There exists a constant $c > 0$ depending on β and ρ such that for any constant $\epsilon > 0$, if $\mathsf{K} \leq csn$ and*

$$q \quad \leq \quad \min\left\{ \left(\frac{k!}{n^\epsilon N_{\max}^k m}\right)^{1/2k}, \left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2}\right)^{1/k}, \left(\frac{k!}{n^\epsilon m^k N_{\max}^k}\right)^{4/k^2} \right\} \tag{6}$$

then we have that $\mathbb{E}[\mathrm{X} \mid \mathrm{F}] = O(n^{-\epsilon})$. Specifically, we can set $\epsilon = 3$ (this choice will be necessary later to apply union bound) and obtain $\mathbb{E}[\mathrm{X} \mid \mathrm{F}] = O(n^{-3})$.

The proof of this lemma is rather technical, and can be found in Appendix A.1.

Furthermore, $|\Lambda| \leq \beta^2$ (which is a constant) and $m = |P_i| \leq |V| = n$. One can then verify that there exist constants c_2^a and c_3^a (which depend on the doubling constant L of μ, the Besicovitch constant β, and the regularity constant ρ), such that if

$$q \leq c_2^a \cdot \left(\frac{1}{n}\right)^{c_3^a/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn}, \tag{7}$$

then the conditions in Eqn. (6) will hold (the simple proof of this can be found in Appendix A.2). Thus, by Lemma 11 and Eqn. (4), we know that if $\mathsf{K} \leq csn$ and (7) holds, then

$$\forall i \in \Lambda, \mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{ has a } uv\text{-clique of size} \geq k+2\right] = O(n^{-3}). \tag{8}$$

On the other hand, note that

$$\mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2|\Lambda|}\right] = \mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{ has a } uv\text{-clique of size} \geq k+2\right]$$

As $|\Lambda|$ is a constant, by Eqn. (3), we obtain that

$$\begin{aligned}&\text{If } \ \forall i \in \Lambda, \mathbb{P}\left[\widehat{G}|_{\tilde{P}_i} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2|\Lambda|}\right] = O(n^{-3}), \ \text{ then} \\ &\qquad\qquad \mathbb{P}\left[\widehat{G}|_{\tilde{A}_{uv}} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2}\right] = O(n^{-3}) \end{aligned} \tag{9}$$

It then follows from Eqn. (8) and (9) that

$$\text{If } \mathsf{K} \le csn \text{ and (7) holds, then } \mathbb{P}\left[\widehat{G}|_{\tilde{A}_{uv}} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] = O(n^{-3}). \quad (10)$$

Set $c_1^a = c \cdot c_2^a \cdot \left(\frac{1}{n}\right)^{c_3^a/(c \ln n)}$. Easy to see that:

$$\text{If } q \le \min\left\{c_1^a,\ c_2^a \cdot \left(\frac{1}{n}\right)^{c_3^a/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn}\right\}, \text{ then}$$
$$\mathbb{P}\left[\widehat{G}|_{\tilde{A}_{uv}} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] = O(n^{-3}). \quad (11)$$

Case (B): bounding the second term in Eqn. (2). Recall that $B_{uv} = \{w \in V \mid w \in B_r^V(u) \cup B_r^V(v)\}$ (see Figure 2 (b)). Imagine we now build the following random graph $\tilde{G}_{uv}^{local} = (\tilde{V}, \tilde{E})$: The vertex set $\tilde{V}$ is simply B_{uv}. To construct the edge set $\tilde{E}$, first, add all edges in the clique spanned by nodes in $B_r^V(u)$ as well as edges in the clique spanned by nodes in $B_r^V(v)$ into $\tilde{E}$. Next, add edge uv to $\tilde{E}$. Finally, insert each crossing edge xy with $x \in B_r^V(u)$ and $y \in B_r^V(v)$ with probability q.

On the other hand, consider the graph $\widehat{G}|_{B_{uv}}$, the induced subgraph of $\widehat{G}$ spanned by vertices in B_{uv}. We can imagine that the graph $\widehat{G}|_{B_{uv}}$ was produced by first taking the induced subgraph $G^*|_{B_{uv}}$, and then insert crossing edges xy each with probability q. Since uv is a bad-edge, by Definition 3, we know that there are no edges between nodes in $B_r^V(u)$ and $B_r^V(v)$ in the random geometric graph G^*. Hence we obtain:

$$\mathbb{P}\left[\widehat{G}|_{B_{uv}} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] \ \le\ \mathbb{P}\left[\tilde{G}_{uv}^{local} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] \quad (12)$$

Using a similar argument as in case (A) (the missing details can be found in Appendix A.3), we have that there exist constants $c_1^b, c_2^b, c_3^b > 0$ which depend on the doubling constant L, the Besicovitch constant β and the regularity constant ρ such that

$$\text{If } q \le \ \min\left\{c_1^b, c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn}\right\}, \text{ then}$$
$$\mathbb{P}\left[\tilde{G}_{uv}^{local} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] = O(n^{-3})$$

Combining this with Eqn. (12), (11) and (2), we know that there exist constants $c_1 = \min\{c_1^a, c_1^b\}$, $c_2 = \min\{c_2^a, c_2^b\}$ and $c_3 = \max\{c_1^a, c_3^b\}$ such that if q satisfies conditions in Eqn. (1), then

$$\mathbb{P}\left[\widehat{G} \text{ has a } uv\text{-clique of size} \ge \mathsf{K}\right] = O(n^{-3})$$

Finally, by applying the union bound, this means:

$$\mathbb{P}\left[\text{for every bad-edge } (u,v),\ \widehat{G} \text{ has a } uv\text{-clique of size} \ge \mathsf{K}\right] = O(n^{-1})$$

Thus with high probability, we have that for every bad-edge (u,v), $\omega_{u,v}(\widehat{G}) < \mathsf{K}$ as long as Eqn. (1) holds. ◀

3.2 Edge clique numbers for the deletion-only case

We now consider the deletion-only case, where we assume that the input graph $\widehat{G} = \widehat{G}^p$ is obtained by deleting each edge in the random geometric graph $G^* = G^*_{\mathcal{X}}(r)$ independently with probability p. For an edge (u, v) in $\widehat{G}$, below we will give a lower bound on the edge-clique number $\omega_{u,v}(\widehat{G})$. A simple observation is that for any edge (u, v) in G^*, as $d_X(u, v) \leq r$, we have that $B_r(u) \cap B_r(v)$ must contain a metric ball of radius $r/2$ (say $B_{r/2}(z)$ centered at midpoint z of a geodesic connecting u to v in X; see Figure 1 (a)). Thus by a similar argument as the proof of Claim 4, the number of points in the $r/2$-ball can be bounded from below w.h.p. Note that all points in a $r/2$-ball span a clique in the random geometric graph G^*. Since we then remove each edge from G^* independently (to obtain $\widehat{G}$), to find a lower bound for $\omega_{u,v}(\widehat{G})$, it suffices to consider the "local" subgraph of $\widehat{G}$ restricted within this $r/2$-ball $B_{r/2}(z)$. This local graph has the same behavior as the standard Erdős–Rényi random graph $G(N_z, 1-p)$, where N_z is the number of points from V within the ball $B_{r/2}(z)$. This eventually leads to the following result.

▶ **Theorem 12.** *Assume Density-cond holds. Let $\widehat{G} = \widehat{G}^p$ denote the final graph after deleting each edge in $G^*_{\mathcal{X}}(r)$ independently with probability p. Then, for any constant $p \in (0, 1)$, with high probability, we have $\omega_{u,v}(\widehat{G}) \geq \frac{2}{3} \log_{1/(1-p)} sn$ for all edges (u, v) in $\widehat{G}$.*

On the high level, we first prove the following technical lemma for an Erdős–Rényi random graph, via an application of Janson's Inequality [1]. The detailed proof can be found in the full version of this paper [15] (see Appendix C in [15]).

▶ **Lemma 13.** *Suppose $G = G(N, \bar{p})$ is an Erdős–Rényi random graph with*

$$\bar{p} \in \left(\left(\frac{1}{N}\right)^{\frac{1}{10}}, \left(\frac{1}{N}\right)^{\frac{1}{64\sqrt[4]{N}}} \right).$$

Set $k := \left\lfloor \log_{1/\bar{p}} N \right\rfloor$. Then, we have

$$\mathbb{P}[\omega(G) < k] < e^{-N^{3/2}}$$

where $\omega(G)$ is the clique number of graph G.

▶ Remark. One can easily verify that $(\frac{1}{N})^{\frac{1}{10}}$ is very close to 0 and $(\frac{1}{N})^{\frac{1}{64\sqrt[4]{N}}}$ is very close to 1 as N goes to infinity. Hence the range $\bar{p} \in \left((\frac{1}{N})^{\frac{1}{10}}, (\frac{1}{N})^{\frac{1}{64\sqrt[4]{N}}} \right)$ is broader (significantly more relaxed) than requiring that $\bar{p}$ is a constant between $(0, 1)$. Hence, while not pursued in the present paper, it is possible to show that Theorem 12 holds for a larger range of p.

Now we are ready to prove the main result in this section (Theorem 12) .

Proof of Theorem 12. Using the argument in the proof of Claim 4, we know that for a fixed good-edge (u, v) (i.e. $d(u, v) \leq r$), with probability $1 - n^{-\frac{8}{3}}$, the geodesic ball $B_{r/2}(z)$ (z is the mid-point of a geodesic connecting u to v in X) contains at least $(sn/4)$ points. Note that all points in a $r/2$-ball form a clique in $r-$neighborhood graph. Since we remove each edge independently, in order to estimate $\omega_{u,v}(\widehat{G})$ from below, it suffices to consider the "local" graph spanned by nodes in this $r/2$-ball. Note that this "local" graph have the same behavior as the standard Erdős–Rényi random graph $G^{loc}_{uv} := G(N_z, 1-p)$, where N_z denotes the number of points falling in the ball $B_{r/2}(z)$.

Furthermore, it is easy to see that, if $N_z \geq sn/4$, then for any *constant* $p \in (0,1)$, one can always find a sufficiently large n such that $1-p \in \left((\frac{1}{N_z})^{\frac{1}{10}}, (\frac{1}{N_z})^{\frac{1}{\sqrt[64]{N_z}}} \right)$.

Now, we are ready to apply Lemma 13 to those "local" graphs: Note that $\bar{p}$ in Lemma 13 will be set to be $1-p$, and N will be set to be N_z.

$$\mathbb{P}\left[\omega(G_{uv}^{loc}) < k \mid N_z \geq \frac{sn}{4}\right] < e^{-(\frac{sn}{4})^{3/2}} \leq e^{-(3\ln n)^{3/2}} < n^{-(\ln n)^{1/2}}$$

where $k = \left\lfloor \log_{1/(1-p)} N_z \right\rfloor$.

Hence for $k' = \frac{2}{3}\log_{1/(1-p)} sn$, we have that

$$\mathbb{P}\left[\omega(G_{uv}^{loc}) < k' \mid N_z \geq \frac{sn}{4}\right] < \mathbb{P}\left[\omega(G_{uv}^{loc}) < k \mid N_z \geq \frac{sn}{4}\right] < n^{-(\ln n)^{1/2}}.$$

By the law of total probability, we know that

$$\begin{aligned}\mathbb{P}\left[\omega(G_{uv}^{loc}) < k' \mid d(u,v) \leq r\right] &< \mathbb{P}\left[\omega(G_{uv}^{loc}) < k' \mid N_z \geq \frac{sn}{4}\right] + \mathbb{P}\left[N_z < \frac{sn}{4}\right] \\ &< n^{-(\ln n)^{1/2}} + n^{-\frac{8}{3}}\end{aligned}$$

Applying the union bound, we have

$$\begin{aligned}\mathbb{P}\left[\bigwedge_{u,v\in V; d(u,v)\leq r} \omega(G_{uv}^{loc}) \geq k'\right] &= 1 - \mathbb{P}\left[\exists u,v \in V \text{ with } d(u,v) \leq r \text{ s.t. } \omega(G_{uv}^{loc}) < k'\right] \\ &\geq 1 - \frac{1}{2}n^2\mathbb{P}\left[\omega(G_{uv}^{loc}) < k' \mid d(u,v) \leq r\right] \\ &\geq 1 - \frac{1}{2}n^{-(\ln n)^{1/2}+2} - \frac{1}{2}n^{-\frac{2}{3}}\end{aligned}$$

Thus, with high probability, for each good-edge (u,v), we have

$$\omega_{u,v}(\widehat{G}) \geq k' = \frac{2}{3}\log_{1/(1-p)} sn.$$ ◀

3.3 Combined Case

In this section, we consider both the deletion and insertion. In other words, we consider the ER-perturbed random geometric graph $\widehat{G}$ generated via the model described in section 2 that includes both edge-deletion probability p and edge-insertion probability q. Our main results for the combined case are summarized in the following theorem. The (somewhat repetitive) details can be found in the full version of this paper [15] (see Appendix D in [15]).

▶ **Theorem 14.** *Assume Assumption-A holds. Let $\widehat{G} = \widehat{G}^{p,q}(r)$ denote the graph obtained by removing each edge in G^* $(= G^*_{\mathcal{X}}(r))$ independently with probability $p \in (0,1)$ and inserting each edge not in G^* independently with probability q. There exist constants $c_1, c_2, c_3 > 0$ which depend on the doubling constant L of μ, the Besicovitch constant $\beta(\mathcal{X})$, and the regularity constant ρ, such that the following holds for any $\mathsf{K} = \mathsf{K}(n)$ with $\mathsf{K} \to \infty$ as $n \to \infty$*

1. *W.h.p., for all good-edges $(u,v) \in \widehat{G}$, $\omega_{u,v}(\widehat{G}) \geq \frac{2}{3}\log_{1/(1-p)} sn$.*
2. *W.h.p., for all bad-edges $(u,v) \in \widehat{G}$, $\omega_{u,v}(\widehat{G}) < \mathsf{K}$ as long as the insertion probability q satisfies*

$$q \leq \min\left\{c_1, \quad c_2 \cdot \left(\frac{1}{n}\right)^{c_3/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn\sqrt{1-p}}\right\} \tag{13}$$

▶ Remark. For example, assume $sn = \Theta(\ln n)$. Then for a constant deletion probability $p \in (0,1)$, w.h.p. the edge clique number for any good-edge is **at least** $\Omega\left(\log_{1/(1-p)} sn\right) = \Omega(\ln \ln n)$. For any bad-edge uv, if the insertion probably $q = o\left((\frac{1}{n})^{\frac{c_3}{\ln \ln n}} \frac{\ln \ln n}{\ln n}\right)$, then its edge clique number is **at most** $\mathsf{K} = o(\ln \ln n)$ w.h.p.. As q decreases, the gap between the edge clique number for good-edges and bad-edges can be made **larger and larger**.

Compared to the insertion-only case, it may seem that the condition on q is too restrictive (recall that for the insertion only case we only require $q = o(1)$ to have a gap between edge clique number for good-edges and bad-edges). Intuitively, this is because: even for an Erdős–Rényi graph $G(n,q)$ with $q = (\frac{1}{n})^{\frac{c_3}{\ln \ln n}} \frac{\ln \ln n}{\ln n}$, its clique number is of order $\Theta(\ln \ln n)$ with high probability[2]. This clique size is already at the same scale as the bound of edge clique number for a good-edge in the deletion-only case. Intuitively, this now gets into a regime where the good/bad-edges potentially have edge cliques of asymptotically similar sizes.

4 Recover the shortest-path metric of $G^*(r)$

In this section, we show an application in recovering the shortest-path metric structure of $G^*_{\mathcal{X}}(r)$ from an input observed graph $\widehat{G}^{p,q}_{\mathcal{X}}(r)$. This problem is previously introduced in [20]. Intuitively, assume that $G^* = G^*_{\mathcal{X}}(r)$ is the true graph of interests (which reflects the metric structure of (X,d)), but the observed graph is a (p,q)-perturbed version $\widehat{G} = \widehat{G}^{p,q}_{\mathcal{X}}(r)$ as described in Section 2. The goal is to recover the shortest-path metric of G^* from its noisy observation $\widehat{G}$ with approximation guarantees. Note that due to the random insertion, two nodes could have significantly shorter path in $\widehat{G}$ than in G^*.

Specifically, given two different metrics defined on the same space (Y, d_1) and (Y, d_2), we say that $d_1 \le \alpha \cdot d_2$ if for any two points $y_1, y_2 \in Y$, we have that $d_1(y_1,y_2) \le \alpha \cdot d_2(y_1,y_2)$. The metric d_1 is an *α-approximation* of d_2 if $\frac{1}{\alpha} \cdot d_2 \le d_1 \le \alpha \cdot d_2$ for $\alpha \ge 1$ and $\alpha = 1$ means that $d_1 = d_2$.

Let d_G denote the shortest-path metric on graph G. It was observed in [20] that, roughly speaking, deletion (with p smaller than a certain constant) does not distort the shortest-path metric of G^* by more than a factor of 2. Insertion however could change shortest-path distances significantly. The authors of [20] then proposed a filtering process to remove some "bad" edges based on the so-called Jaccard index, and showed that after the Jaccard-filtering process, the shortest-path metric of the resulting graph $\tilde{G}$ 2-approximates that of the true graph G^* when the insertion probability q is small.

We follow the same framework as [20], but change the filtering process to be one based on the edge clique number instead. This allows us to recover the shortest-path metric within constant factor for a much larger range of values of the insertion probability q, although we do need the extra Regularity-cond which is not needed in [20]. (Note that it does not seem that the bound of [20] can be improved even with this extra Regularity-cond).

We now introduce our edge-clique based filtering process.

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

The following result can be proved by almost the same argument as that for Theorem 12 of [20] with the help of Theorem 14.

[2] Indeed, the upper bound can be easily derived by computing the expectation; and Lemma 13 in the Appendix provides the lower bound.

▶ **Theorem 15.** *Assume Assumption-A holds. Suppose $\widehat{G} = \widehat{G}^{p,q}(r)$ is the graph as in Theorem 14. Let $\tilde{G}_\tau$ denote the resulting graph after τ-Clique filtering. Then there exist constants $c_0, c_1, c_2, c_3 > 0$ which depend on the doubling constant L of μ, the Besicovitch constant $\beta(\mathcal{X})$, and the regularity constant ρ, such that if $p \in (0, c_0)$, $\tau \leq \frac{2}{3}\log_{1/(1-p)} sn$, and*

$$q \leq c_2 \cdot \left(\frac{1}{n}\right)^{c_3/\tau} \cdot \frac{\tau}{sn\sqrt{1-p}} \qquad \left(= \min\left\{\ c_1, \quad c_2 \cdot \left(\frac{1}{n}\right)^{c_3/\tau} \cdot \frac{\tau}{sn\sqrt{1-p}}\ \right\}\right),$$

then, with high probability, the shortest-path metric $d_{\tilde{G}_\tau}$ is a 3-approximation of the shortest-path metric d_{G^*} of G^*. However, if the deletion probability $p = 0$, then we have w.h.p. that $d_{\tilde{G}_\tau}$ is a 3-approximation of d_{G^*} as long as $\tau < \frac{sn}{4}$, and $q \leq \min\left\{\ c_1, \quad c_2 \cdot \left(\frac{1}{n}\right)^{c_3/\tau} \cdot \frac{\tau}{sn}\ \right\}$.

Proof. A simple application of Theorem 14 (i) and (ii) gives the following two lemmas, respectively.

▶ **Lemma 16.** *Under the same setting as Theorem 14, if $p \in (0, 1)$ and the filtering parameter τ satisfies $\tau < \frac{2}{3}\log_{1/(1-p)} sn$, then, with high probability, our τ-Clique filtering process will not remove any good-edges.*

▶ **Lemma 17.** *Under the same setting as Theorem 14, there exist constants $c_1, c_2, c_3 > 0$ such that for constant $p \in (0, 1)$, with high probability, a τ-Clique filtering process deletes all bad-edges, as long as $q \leq \min\left\{c_1, \quad c_2 \cdot (\frac{1}{n})^{c_3/\tau} \cdot \frac{\tau}{sn\sqrt{1-p}}\ \right\}$.*

Our goal is to show that $\frac{1}{3}d_{\tilde{G}_\tau} \leq d_{G^*} \leq 3d_{\tilde{G}_\tau}$. Let $\mathcal{E}_1$ denote the event where $d_{\widehat{G}\cap G^*} \leq 2d_{G^*}$. By Lemma 17 of [20], event $\mathcal{E}_1$ happens with probability at least $1 - n^{-\Omega(1)}$.

Let $\mathcal{E}_2$ denote the event where all edges $\widehat{G} \cap G^*$ are also contained in the edge set of the filtered graph $\tilde{G}_\tau$; that is, $\widehat{G} \cap G^* \subseteq \tilde{G}_\tau$. By Lemma 16, event $\mathcal{E}_2$ happens with probability at least $1 - n^{-\frac{2}{3}}$ (this bound is derived in the proof of Theorem 12). It then follows that:

If both events $\mathcal{E}_1$ and $\mathcal{E}_2$ happen, then $d_{\tilde{G}_\tau} \leq d_{\widehat{G}\cap G^*} \leq 2d_{G^*} \leq 3d_{G^*}$.

What remains is to show $d_{G^*} \leq 3d_{\tilde{G}_\tau}$. To this end, we define $\mathcal{E}_3$ to be the event where for all bad-edges (u, v) in $\widehat{G}$, we have $\omega_{u,v}(\widehat{G}) < \tau$. If $\mathcal{E}_3$ happens, then it implies that for an arbitrary edge $(u, v) \in E(\tilde{G}_\tau)$, either $(u, v) \in E(G^*)$ (thus $d_{G^*}(u, v) = 1$) or $d_{G^*}(u, v) \leq 3$ (since there is at least one edge connecting $N_{G^*}(u)$ and $N_{G^*}(v)$). By Lemma 17, event $\mathcal{E}_3$ happens with probability at least $1 - o(1)$ (the exact bound can be found in the proof of Theorem 14).

By applying the union bound, we know that $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ happen simultaneously with high probability.

Using a similar argument as the proof of Theorem 11 in [20], it then follows that given any $u, v \in V$ connected in $\tilde{G}_\tau$, we can find a path in G^* of at most $3d_{\tilde{G}_\tau}(u, v)$ number of edges to connect u and v. Furthermore, event $\mathcal{E}_1$ implies that if u and v are not connected in $\tilde{G}_\tau$, then they cannot be connected in G^* either. Putting everything together, we thus obtain $d_{G^*} \leq 3d_{\tilde{G}_\tau}$. Theorem 15 then follows. ◀

▶ Remark. Consider the insertion-only case (i.e, the deletion probability $p = 0$), which is a case of independent interest. In this case, if we choose $\tau = \ln n$ and assume that $sn > 4\tau$, then w.h.p. 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 $sn > 4\tau = 4\ln n$), then q is only required to be smaller than a constant. If $sn = \ln^a n$ for some $a > 1$, then we require that $q \leq \frac{c}{\ln^{a-1} n}$. In constrast, [20] requires that $q = o(s)$, which is $q = o(\frac{\ln^c n}{n})$ if $sn = \ln^a n$ with $a \geq 1$. The gap (ratio) between these two bounds is nearly a factor of n.

For a *constant* deletion probability $p \in (0, c_0)$, our clique filtering process still requires a much larger range of insertion probability q compared to what's required in [20]. For example, assume $sn = \Theta(\ln n)$. Then if we choose the filtering parameter to be $\tau = \sqrt{\ln \ln n}$, then we can recover d_{G^*} approximately as long as the insertion probability $q = o\left(\left(\frac{1}{n}\right)^{\frac{c_3}{\sqrt{\ln \ln n}}} \frac{\sqrt{\ln \ln n}}{\ln n}\right)$. This is still much larger than the q required in [20], which is $q = o(s) = o\left(\frac{\ln n}{n}\right)$. In fact, $\left(\frac{1}{n}\right)^{\frac{c_3}{\sqrt{\ln \ln n}}} \frac{\sqrt{\ln \ln n}}{\ln n}$ is asymptoticly larger than $\frac{1}{n^{\varepsilon}}$ for any $\varepsilon > 0$. However, we do point out that the Jaccard-filtering process in [20] 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.

References

1 Noga Alon and Joel Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.

2 Philippe Blanchard and Dimitri Volchenkov. *Mathematical analysis of urban spatial networks*. Springer Science & Business Media, 2008.

3 Béla Bollobás and Oliver Riordan. *Percolation*. Cambridge University Press, 2006.

4 Lorna Booth, Jehoshua Bruck, Matthew Cook, and Massimo Franceschetti. Ad hoc wireless networks with noisy links. In *Proceedings of IEEE International Symposium on Information Theory*, pages 386–386. IEEE, 2003.

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

6 Don Coppersmith, David Gamarnik, and Maxim Sviridenko. *The Diameter of a Long-Range Percolation Graph*, pages 147–159. Birkhäuser Basel, Basel, 2002.

7 Carl Dettmann and Orestis Georgiou. Random geometric graphs with general connection functions. *Physical Review E*, 93(3):032313, 2016.

8 Herbert Federer. *Geometric measure theory*. Springer, 2014.

9 Janos Galambos. Bonferroni inequalities. *The Annals of Probability*, pages 577–581, 1977.

10 Edward Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

11 Piyush Gupta and Panganamala Kumar. Critical power for asymptotic connectivity in wireless networks. In *Stochastic analysis, control, optimization and applications*, pages 547–566. Springer, 1999.

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

13 Svante Janson, Róbert Kozma, Miklós Ruszinkó, and Yury Sokolov. Bootstrap percolation on a random graph coupled with a lattice. *Electronic Journal of Combinatorics*, 2016.

14 Antti Kaenmaki, Tapio Rajala, and Ville Suomala. Local homogeneity and dimensions of measures. *Annali Della Scoula Normale Superiore Di Pisa-Classe Di Scienze*, 16(4):1315–1351, 2016.

15 Matthew Kahle, Minghao Tian, and Yusu Wang. Local cliques in ER-perturbed random geometric graphs. *arXiv preprint*, 2018. `arXiv:1810.08383`.

16 Colin McDiarmid and Tobias Müller. On the chromatic number of random geometric graphs. *Combinatorica*, 31(4):423–488, 2011.

17 Ronald Meester and Rahul Roy. *Continuum percolation*, volume 119. Cambridge University Press, 1996.

18 Maziar Nekovee. Worm epidemics in wireless ad hoc networks. *New Journal of Physics*, 9(6):189, 2007.

19 Mark Newman. Random graphs as models of networks. *Handbook of Graphs and Networks: From the Genome to the Internet*, pages 35–68, 2002.

20 Srinivasan Parthasarathy, David Sivakoff, Minghao Tian, and Yusu Wang. A Quest to Unravel the Metric Structure Behind Perturbed Networks. In *33rd International Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia*, pages 53:1–53:16, 2017. doi:10.4230/LIPIcs.SoCG.2017.53.

21 Mathew Penrose. The longest edge of the random minimal spanning tree. *Ann. Appl. Probab.*, 7(2):340–361, May 1997.

22 Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

23 Gareth Peters and Tomoko Matsui. *Theoretical Aspects of Spatial-Temporal Modeling.* Springer, 2015.

24 Xian Yuan Wu. Mixing Time of Random Walk on Poisson Geometry Small World. *Internet Mathematics*, 2017.

A The missing proofs in Section 3.1

A.1 The proof of Lemma 11

We first pick $c(\beta)$ to be a positive constant such that $\left\lfloor \frac{c(\beta)N_{\max}}{2|\Lambda|} \right\rfloor - 2 \le N_{\max}$. Then, since $k = \left\lfloor \frac{\mathsf{K}}{2|\Lambda|} \right\rfloor - 2$, it is easy to see that for any $\mathsf{K} \le c(\beta)N_{\max} = c(\beta)3\rho s n$, we have $k \le N_{\max}$. Pick $c = c(\beta)3\rho$. Then, $\mathsf{K} \le csn$ implies $k \le N_{\max}$.

To estimate the summation on the right hand side of Eqn. (5), we consider the quantity $x_{\max} := \max_i \{x_i\}$. We first enumerate all the possible cases of $(x_1, x_2, \cdots, x_m)$ when $x_{\max}$ is fixed, and then vary the value of $x_{\max}$.

Set $h(y) = \max_{x_{\max}=y} \left\{ \sum_{i=1}^m x_i^2 \right\}$ for $y \ge \lceil \frac{k}{m} \rceil$. It is the maximum value of $\sum_{i=1}^m x_i^2$ under the constraint $x_{\max} = y$. Without loss of generality, we assume $x_1 = y$ and $y \ge x_2 \ge x_3 \ge \cdots \ge x_m \ge 0$. We argue that $\arg\max_{x_{\max}=y} \left\{ \sum_{i=1}^m x_i^2 \right\} = \{y, y, \cdots, y, k - ry, 0, \cdots, 0\}$, that is $x_1 = x_2 = \cdots = x_r = y, x_{r+1} = k - ry$ where $r = \left\lfloor \frac{k}{y} \right\rfloor$.

To show this, we first consider x_2: if $x_2 = y$, then consider x_3; otherwise, $x_2 < y$, then we search for the largest index j such that $x_j > 0$. Note the fact that if $x \ge y > 0$, then $(x+1)^2 + (y-1)^2 = x^2 + y^2 + 2(x-y) + 2 > x^2 + y^2$. So if we increase x_2 by 1 and decrease x_j by 1, we will enlarge $\sum_{i=1}^m x_i^2$. After we update $x_2 = x_2 + 1$, $x_j = x_j - 1$, we still get a decreasing sequence $x_1 \ge x_2 \ge \cdots \ge x_m \ge 0$. If we still have $x_2 < y$, then we repeat the same procedure above (by increasing x_2 and decreasing x_j where j is the largest index such that $x_j > 0$). We repeat this process until $x_2 = y$ or $x_1 + x_2 = k$. If it is the former case (i.e, $x_2 = y$), then we consider x_3 and so on. Finally, we will get the sequence $x_1 = \cdots = x_r = y, x_{r+1} = k - ry$ where $r = \left\lfloor \frac{k}{y} \right\rfloor$ as claimed, and this maximizes $\sum_{i=1}^m x_i^2$.

Next we claim that $h(y+1) > h(y)$. The reason is similar to the above. We update the sequence $x_1 = x_2 = \cdots = x_r = y, x_{r+1} = k - ry$ (which corresponding to $h(y)$) from x_1: we increase x_1 by 1; search the largest index s such that $x_s > 0$ and decrease x_s by 1. And then consider x_2 and so on and so forth. This process won't stop until $x_1 = x_2 = \cdots = x_q = y+1$ and $x_{q+1} = k - q(y+1)$ with $q = \left\lfloor \frac{k}{y+1} \right\rfloor$. Thus $h(y+1) > h(y)$.

By enumerating all the possible values of $x_{\max}$, we split Eqn. (5) into three parts as follows (corresponding to $x_{\max} = k, x_{\max} \in \left[\lceil \frac{k+1}{2} \rceil, k-1\right]$ and $x_{\max} \in \left[\lceil \frac{k}{m} \rceil, \lceil \frac{k+1}{2} \rceil - 1\right]$) (see the remarks after this equation for how the inequality is derived);

$$
\begin{aligned}
&q^{2k} \sum_{\substack{x_1+x_2+\cdots+x_m=k \\ x_i \geq 0}} \binom{N_{\max}}{x_1}\binom{N_{\max}}{x_2}\cdots\binom{N_{\max}}{x_m} q^{(k^2-\sum_{i=1}^m x_i^2)/2} \\
&\leq q^{2k}\binom{N_{\max}}{k} m \;+\; q^{2k} \sum_{x_{\max}=\lceil\frac{k+1}{2}\rceil}^{k-1} \Bigg(\binom{m}{1}\binom{N_{\max}}{x_{\max}} \sum_{\substack{y_1+\cdots+y_{m-1}=k-x_{\max} \\ x_{\max}\geq y_i \geq 0}} \binom{N_{\max}}{y_1}\cdots \\
&\binom{N_{\max}}{y_{m-1}} q^{x_{\max}(k-x_{\max})} \Bigg) \;+\; \binom{mN_{\max}}{k} q^{\frac{(k-1)^2}{4}+2k}.
\end{aligned} \tag{14}
$$

▶ Remark. The first term on the right hand side of Eqn. (14) comes from the fact that if $x_{\max} = k$, then there are m possible cases for $(x_1, x_2, \cdots, x_m)$. For each case, the value of each term in the summation is $\binom{N_{\max}}{k}$, giving rise to the first term in Eqn. (14).

The third term on the right hand side of Eqn. (14) can be derived as follows. First, observe that

$$
\begin{aligned}
&\sum_{x_{\max}=\lceil\frac{k}{m}\rceil}^{\lceil\frac{k+1}{2}\rceil-1} \Bigg(\sum_{\substack{x_1+x_2+\cdots+x_m=k \\ x_i\geq 0, \max_i\{x_i\}=x_{\max}}} \binom{N_{\max}}{x_1}\binom{N_{\max}}{x_2}\cdots\binom{N_{\max}}{x_m} \Bigg) \\
&\leq \sum_{\substack{x_1+x_2+\cdots+x_m=k \\ x_i\geq 0}} \binom{N_{\max}}{x_1}\binom{N_{\max}}{x_2}\cdots\binom{N_{\max}}{x_m} = \binom{mN_{\max}}{k}.
\end{aligned}
$$

On the other hand, as $x_{\max} \leq \lceil\frac{k+1}{2}\rceil - 1 = \lceil\frac{k-1}{2}\rceil$, we have:

$$
\frac{k^2 - \sum_{i=1}^m x_i^2}{2} \geq \frac{k^2 - h(x_{\max})}{2} \geq \frac{k^2 - h(\lceil\frac{k-1}{2}\rceil)}{2} \geq \frac{(k-1)^2}{4},
$$

where the second inequality uses the fact that $h(y)$ is an increasing function, and the last inequality comes from that $h(\lceil\frac{k-1}{2}\rceil) \leq (\lceil\frac{k-1}{2}\rceil)^2 + (\lceil\frac{k-1}{2}\rceil)^2 + 1 \leq k^2/4 + k^2/4 + 1 = k^2/2 + 1$.

To this end, it suffices to estimate all three terms on the right hand side of Eqn. (14).

The first term of Eqn. (14): According to the assumptions in Eqn. (6), we know

$$
q \leq \left(\frac{k!}{n^{\epsilon} N_{\max}^k m}\right)^{1/2k}.
$$

Thus, for the first term of Eqn. (14), we have:

$$
q^{2k}\binom{N_{\max}}{k} m \;\leq\; \left(\frac{k!}{n^{\epsilon} N_{\max}^k m}\right) \frac{N_{\max}^k}{k!} m \;=\; \frac{1}{n^{\epsilon}}. \tag{15}
$$

The second term of Eqn. (14): For the second term of Eqn. (14), we relax the constraint $x_{\max} \geq y_i \geq 0$ to $y_i \geq 0$. Thus, we have:

$$
\begin{aligned}
\sum_{\substack{y_1+\cdots+y_{m-1}=k-x_{\max} \\ x_{\max} \;\geq\; y_i \geq 0}} \binom{N_{\max}}{y_1}\cdots\binom{N_{\max}}{y_{m-1}} &\leq \sum_{\substack{y_1+\cdots+y_{m-1}=k-x_{\max} \\ y_i\geq 0}} \binom{N_{\max}}{y_1}\cdots\binom{N_{\max}}{y_{m-1}} \\
&= \binom{(m-1)N_{\max}}{k-x_{\max}} \;\leq\; \frac{(m-1)^{k-x_{\max}} N_{\max}^{k-x_{\max}}}{(k-x_{\max})!}.
\end{aligned} \tag{16}
$$

Now apply (16) to the second term of (14), we have (starting from the second line, we replace x_{max} to be j for simplicity):

$$
\begin{aligned}
& q^{2k} \sum_{x_{\max}=\lceil \frac{k+1}{2} \rceil}^{k-1} \left(\binom{m}{1} \binom{N_{\max}}{x_{\max}} \sum_{\substack{y_1+\cdots+y_{m-1}=k-x_{\max} \\ x_{\max} \geq y_i \geq 0}} \binom{N_{\max}}{y_1} \cdots \binom{N_{\max}}{y_{m-1}} q^{x_{\max}(k-x_{\max})} \right) \\
\leq & \sum_{j=\lceil \frac{k+1}{2} \rceil}^{k-1} \left(m \frac{(N_{\max})^j}{j!} q^{2k+j(k-j)} \frac{(m-1)^{k-j} N_{\max}^{k-j}}{(k-j)!} \right) \\
< & \sum_{j=\lceil \frac{k+1}{2} \rceil}^{k-1} \left(m^{k-j+1} N_{\max}^k \binom{k}{k-j} \frac{1}{k!} q^{2k+j(k-j)} \right) \\
< & \sum_{j=\lceil \frac{k+1}{2} \rceil}^{k-1} \left(m^{k-j+1} N_{\max}^k \frac{k^{k-j}}{(k-j)!} \frac{1}{k!} q^{2k+j(k-j)} \right). \qquad (17)
\end{aligned}
$$

Since $q \leq \left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} \right)^{1/k}$ by Eqn. (6), for each j satisfying $\lceil \frac{k+1}{2} \rceil \leq j \leq k-1$, we have:

$$
\begin{aligned}
& m^{k-j+1} N_{\max}^k \frac{k^{k-j}}{(k-j)!} \frac{1}{k!} q^{2k+j(k-j)} \\
\leq & m^{k-j+1} N_{\max}^k \frac{k^{k-j}}{(k-j)!} \frac{1}{k!} \frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} \left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} \right)^{\frac{2k+j(k-j)}{k}-1} \\
\leq & m^{k-j-1} k^{k-j-1} \left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} \right)^{\frac{k+j(k-j)}{k}} \frac{1}{kn^\epsilon} \\
\leq & m^{k-j-1} k^{k-j-1} \left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} \right)^{\frac{k-j-1}{2}} \frac{1}{kn^\epsilon} \qquad (18) \\
= & \left(\frac{k!}{n^\epsilon N_{\max}^k} \right)^{\frac{k-j-1}{2}} \frac{1}{kn^\epsilon} \\
\leq & \frac{1}{kn^\epsilon}. \qquad (19)
\end{aligned}
$$

Eqn. (18) comes from two facts: 1) $k \leq N_{\max}$ (and thus the term $\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2} < 1$) and 2) by tedious by elementary calculation, we can show that $\frac{k+j(k-j)}{k} \geq \frac{k-j-1}{2}$ when $\lceil \frac{k+1}{2} \rceil \leq j \leq k-1$. Eqn. (19) holds since $k \leq N_{\max}$ (and thus $\frac{k!}{n^\epsilon N_{\max}^k} < 1$).

The third term of Eqn. (14): For the third term of (14), plugging in the condition

$$
q \leq \left(\frac{k!}{n^\epsilon m^k N_{\max}^k} \right)^{\frac{4}{k^2}},
$$

we thus have

$$\binom{mN_{\max}}{k} q^{\frac{(k-1)^2}{4}+2k} \leq \frac{(mN_{\max})^k}{k!} q^{\frac{(k-1)^2}{4}+2k}$$
$$\leq \frac{(mN_{\max})^k}{k!} \frac{k!}{n^\epsilon m^k N_{\max}^k} \left(\frac{k!}{n^\epsilon m^k N_{\max}^k}\right)^{\frac{4}{k^2}\left(\frac{(k-1)^2}{4}+2k\right)-1} \leq \frac{1}{n^\epsilon} \tag{20}$$

where the last inequality holds as $\frac{k!}{n^\epsilon m^k N_{\max}^k} < 1$.

Finally, combining (15), (19) and (20), we have:

$$E[\mathrm{X} \mid \mathrm{F}] \leq \frac{1}{n^\epsilon} + \frac{k}{2} \cdot \frac{1}{kn^\epsilon} + \frac{1}{n^\epsilon} = \frac{5}{2n^\epsilon}.$$

This proves Lemma 11.

A.2 Existences of constants c_2^a and c_3^a

We claim that there exist constants c_2^a and c_3^a (which depend on the doubling constant L of μ, the Besicovitch constant β, and the regularity constant ρ), such that if

$$q \leq c_2^a \cdot \left(\frac{1}{n}\right)^{c_3^a/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn},$$

then the conditions in Eqn. (6) will hold. We prove this by elementary calculation below, where we will use the Stirling's approximation $k! > \sqrt{2\pi} k^k e^{-k}$ and the fact that $k \leq N_{\max} = 3\rho sn$ ($\mathsf{K} \leq csn$ implies this due to the choice of c in the proof of Lemma 11) and $m \leq n$ (where recall that m is the size of the number of clusters in the clique-decomposition of P_i).

$$\left(\frac{k!}{n^\epsilon N_{\max}^k m}\right)^{\frac{1}{2k}} > \left(\frac{\sqrt{2\pi}k^k e^{-k}}{n^{1+\epsilon}}\right)^{\frac{1}{2k}} \left(\frac{1}{3\rho sn}\right)^{\frac{1}{2}} = \left(e^{-\frac{1}{2}}(2\pi)^{\frac{1}{4k}}\right) \left(\frac{1}{n}\right)^{\frac{1+\epsilon}{2k}} \left(\frac{k}{3\rho sn}\right)^{\frac{1}{2}}$$
$$\left(\frac{k!}{k^2 n^\epsilon N_{\max}^k m^2}\right)^{\frac{1}{k}} > \left(e^{-1}(2\pi)^{\frac{1}{4k}} k^{-\frac{2}{k}}\right) \left(\frac{1}{n}\right)^{\frac{2+\epsilon}{k}} \left(\frac{k}{3\rho sn}\right)$$
$$\left(\frac{k!}{n^\epsilon m^k N_{\max}^k}\right)^{\frac{4}{k^2}} > \left((2\pi)^{\frac{2}{k^2}} e^{-\frac{4}{k}}\right) \left(\frac{1}{n}\right)^{\frac{4(k+\epsilon)}{k^2}} \left(\frac{k}{3\rho sn}\right)^{\frac{4}{k}}.$$

Thus, by comparing the exponents of each term, we know that if $q \leq \frac{1}{3\rho e}\left(\frac{1}{n}\right)^{\frac{4+\epsilon}{k}}\left(\frac{k}{sn}\right)$, then the condition on q holds. Finally, since $k = \left\lfloor \frac{\mathsf{K}}{2|\Lambda|} \right\rfloor - 2$, easy to see there exists constants c_2^a, c_3^a such that the constraint $q \leq c_2^a \left(\frac{1}{n}\right)^{c_3^a/\mathsf{K}} \frac{\mathsf{K}}{sn}$ implies the condition.

A.3 The missing details in case (B) of Theorem 10

Denote by H the event that "for every $v \in V$, the ball $B_r(v) \cap V$ contains at most $3L\rho sn$ points", and H^c is its complement. By an argument similar to that of Claim 9, we have that $\mathbb{P}[\mathrm{H}^c] \leq n^{-5}$. Set $N_u := |B_r^V(u)|$ and $N_v := |B_r^V(v)|$. Let $\tilde{k} := \left\lfloor \frac{\mathsf{K}}{2} \right\rfloor - 2$. For every set S of $(\tilde{k}+2)$ vertices in $\tilde{G}_{uv}^{local}$, let A_S be the event "*S is a uv-clique in $\tilde{G}_{uv}^{local}$*" and Y_S its indicator random variable. Set

$$\mathrm{Y} = \sum_{|S|=\tilde{k}+2} \mathrm{Y}_S.$$

Then Y is the number of uv-cliques of size $(\tilde{k}+2)$ in $\tilde{G}^{local}_{uv}$. Linearity of expectation gives:

$$\mathbb{E}[\mathrm{Y} \mid \mathrm{H}] = \sum_{|S|=\tilde{k}+2} \mathbb{E}[\mathrm{Y}_S \mid \mathrm{H}] = \sum_{\substack{x_1+x_2=\tilde{k} \\ 0\le x_1\le N_u-1 \\ 0\le x_2\le N_v-1}} \binom{N_u-1}{x_1}\binom{N_v-1}{x_2} q^{(x_1+1)(x_2+1)-1}. \tag{21}$$

To estimate this quantity, we first prove the following result:

▶ **Lemma 18.** *For any constant $\epsilon > 0$, we have that $\mathbb{E}[\mathrm{Y} \mid \mathrm{H}] = O(n^{-\epsilon})$ as long as the following condition on q holds:*

$$q \;\le\; \min\left\{ \left(\frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon}(N_u+N_v)^{\tilde{k}}}\right)^{1/\tilde{k}}, \;\; \left(\frac{\tilde{k}!}{n^{\epsilon}(N_u+N_v)^{\tilde{k}}}\right)^{16/\tilde{k}^2} \right\}. \tag{22}$$

Specifically, setting $\epsilon = 3$ (a case which we will use later), we have $\mathbb{E}[\mathrm{Y} \mid \mathrm{H}] = O(n^{-3})$.

The proof of this technical result can be found in Appendix A.4.

Note that if event H is true, then $N_u + N_v \le 6L\rho sn$. In this case, there exist two constants c_2^b and c_3^b which depend on the doubling constant L of μ, the Besicovitch constant β, and the regularity constant ρ, such that if $\mathsf{K} \le 12L\rho sn$ and

$$q \le c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn},$$

then the conditions in Eqn. (22) will hold (the simple proof of this can be found in Appendix A.5).

On the other hand, we have

$$\begin{aligned}
\mathbb{P}\left[\tilde{G}^{local}_{uv} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] &= \mathbb{P}[\mathrm{Y}>0] \\
&= \mathbb{P}[\mathrm{Y}>0 \mid \mathrm{H}]\cdot\mathbb{P}[\mathrm{H}] + \mathbb{P}[\mathrm{Y}>0 \mid \mathrm{H}^c]\cdot\mathbb{P}[\mathrm{H}^c] \\
&\le \mathbb{P}[\mathrm{Y}>0 \mid \mathrm{H}] + \mathbb{P}[\mathrm{H}^c] \\
&\le \mathbb{E}[\mathrm{Y} \mid \mathrm{H}] + n^{-5}.
\end{aligned}$$

Thus, by Lemma (18), we know that

$$\text{If } \mathsf{K} \le 6L\rho sn \text{ and } q \le c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn}, \text{ then}$$
$$\mathbb{P}\left[\tilde{G}^{local}_{uv} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] = O(n^{-3}). \tag{23}$$

Finally, suppose $\mathsf{K} > \mathsf{K}_1 = 12L\rho sn$. Set

$$c_1^b = c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/(12L\rho \ln n)} \cdot \frac{\mathsf{K}_1}{sn} \le c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/\mathsf{K}_1} \cdot \frac{\mathsf{K}_1}{sn}$$

where the inequality holds as by Assumption-A $sn > \ln n$. Plugging in $\mathsf{K}_1 = 12L\rho sn$ to the definition of c_1^b, it is then easy to see that c_1^b is a positive constant. Using Eqn. (23), we know that if $q \le c_1^b$ and $\mathsf{K} > \mathsf{K}_1 = 12L\rho sn$, then

$$\begin{aligned}
\mathbb{P}\left[\tilde{G}^{local}_{uv} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}}{2}\right] &\le \mathbb{P}\left[\tilde{G}^{local}_{uv} \text{ has a } uv\text{-clique of size} \ge \frac{\mathsf{K}_1}{2}\right] \\
&= O(n^{-3}).
\end{aligned}$$

Combining this with the discussion above and applying Lemma 18, we have

$$\text{If } q \leq \min\left\{c_1^b, c_2^b \cdot \left(\frac{1}{n}\right)^{c_3^b/\mathsf{K}} \cdot \frac{\mathsf{K}}{sn}\right\}, \text{ then}$$
$$\mathbb{P}\left[\tilde{G}_{uv}^{local} \text{ has a } uv\text{-clique of size} \geq \frac{\mathsf{K}}{2}\right] = O(n^{-3}).$$

A.4 The proof of Lemma 18

Proof. It is easy to see that if $\mathsf{K} > 2(N_u + N_v)$, then $\tilde{k} > N_u + N_v - 2$ which implies that the summation on the right hand side of Eqn. (21) is 0. Now let's focus on the case when $\mathsf{K} \leq 2(N_u + N_v)$. In this case, we have $\tilde{k} \leq (N_u - 1) + (N_v - 1) < N_u + N_v$. Note that the right hand side of (21) can be bounded from above by:

$$\sum_{\substack{x_1+x_2=\tilde{k}\\ 0\leq x_1\leq N_u-1\\ 0\leq x_2\leq N_v-1}} \binom{N_u-1}{x_1}\binom{N_v-1}{x_2} q^{(x_1+1)(x_2+1)-1} \;\leq\; q^{\tilde{k}} \sum_{i=0}^{\tilde{k}} \binom{N_u}{i}\binom{N_v}{\tilde{k}-i} q^{i(\tilde{k}-i)}$$
$$\leq q^{\tilde{k}} \left(\sum_{i=0}^{\lfloor \frac{\tilde{k}}{4} \rfloor} \left[\binom{N_u}{i}\binom{N_v}{\tilde{k}-i} + \binom{N_u}{\tilde{k}-i}\binom{N_v}{i} \right] q^{i(\tilde{k}-i)} \right) + \binom{N_u+N_v}{\tilde{k}} q^{\tilde{k}+\frac{\tilde{k}^2}{16}}. \tag{24}$$

Eqn. (24) is due to the fact that when $\left\lfloor \frac{\tilde{k}}{4} \right\rfloor + 1 \leq i \leq \tilde{k} - \left\lfloor \frac{\tilde{k}}{4} \right\rfloor - 1$, we have

$$i(\tilde{k}-i) \geq \left(\left\lfloor \frac{\tilde{k}}{4} \right\rfloor + 1\right)\left(\left\lfloor \frac{\tilde{k}}{4} \right\rfloor + 1\right) \geq \frac{\tilde{k}^2}{16}.$$

Now it suffices to estimate the two terms on the right hand side of Eqn. (24).

The first term of Eqn. (24): For the first term of (24), we have the following estimate:

$$\left[\binom{N_u}{i}\binom{N_v}{\tilde{k}-i} + \binom{N_u}{\tilde{k}-i}\binom{N_v}{i}\right] q^{\tilde{k}+i(\tilde{k}-i)} \leq \frac{N_u^i N_v^{\tilde{k}-i} + N_u^{\tilde{k}-i} N_v^i}{i!(\tilde{k}-i)!} q^{\tilde{k}} q^{i(\tilde{k}-i)}$$
$$\leq \frac{(N_u+N_v)^{\tilde{k}}}{i!(\tilde{k}-i)!} q^{\tilde{k}} q^{i(\tilde{k}-i)}.$$

By plugging in the condition $q \leq \left(\frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon} (N_u+N_v)^{\tilde{k}}}\right)^{1/\tilde{k}}$, we have:

$$\frac{(N_u+N_v)^{\tilde{k}}}{i!(\tilde{k}-i)!} q^{\tilde{k}} q^{i(\tilde{k}-i)} \leq \frac{(N_u+N_v)^{\tilde{k}}}{i!(\tilde{k}-i)!} \frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon}(N_u+N_v)^{\tilde{k}}} q^{i(\tilde{k}-i)} = \frac{\tilde{k}!}{i!(\tilde{k}-i)!} q^{i(\tilde{k}-i)} \frac{1}{\tilde{k}^2 n^{\epsilon}}.$$

For $i = 0$, we have $\frac{\tilde{k}!}{i!(\tilde{k}-i)!} q^{i(\tilde{k}-i)} \frac{1}{\tilde{k}^2 n^{\epsilon}} = \frac{1}{\tilde{k}^2 n^{\epsilon}}$. For $i \geq 1$, note that $1 \leq i \leq \left\lfloor \frac{\tilde{k}}{4} \right\rfloor$ implies

$\frac{i(\tilde{k}-i)}{\tilde{k}} \geq \frac{i}{2}$. Thus, we have:

$$\begin{aligned}
\frac{\tilde{k}!}{i!(\tilde{k}-i)!} q^{i(\tilde{k}-i)} \frac{1}{\tilde{k}^2 n^{\epsilon}} &\leq \frac{\tilde{k}^i}{i!} \left(\frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon} (N_u + N_v)^{\tilde{k}}} \right)^{\frac{i(\tilde{k}-i)}{\tilde{k}}} \frac{1}{\tilde{k}^2 n^{\epsilon}} \\
&\leq \frac{\tilde{k}^i}{i!} \left(\frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon} (N_u + N_v)^{\tilde{k}}} \right)^{\frac{i}{2}} \frac{1}{\tilde{k}^2 n^{\epsilon}} \\
&\leq \left(\frac{\tilde{k}!}{n^{\epsilon} (N_u + N_v)^{\tilde{k}}} \right)^{\frac{i}{2}} \frac{1}{\tilde{k}^2 n^{\epsilon}} \\
&\leq \frac{1}{\tilde{k}^2 n^{\epsilon}}.
\end{aligned}$$

The last two inequalities hold since $\tilde{k} \leq N_u + N_v$. Therefore,

$$q^{\tilde{k}} \sum_{i=0}^{\lfloor \frac{\tilde{k}}{4} \rfloor} \left[\binom{N_u}{i} \binom{N_v}{\tilde{k}-i} + \binom{N_u}{\tilde{k}-i} \binom{N_v}{i} \right] q^{i(\tilde{k}-i)} \leq \frac{\tilde{k}}{4} \frac{1}{\tilde{k}^2 n^{\epsilon}} = \frac{1}{4\tilde{k} n^{\epsilon}}. \tag{25}$$

The second term of Eqn. (24): For the second term of (24), directly plugging in the condition $q \leq \left(\frac{\tilde{k}!}{(N_u+N_v)^{\tilde{k}} n^{\epsilon}} \right)^{16/\tilde{k}^2}$, we have:

$$\binom{N_u + N_v}{\tilde{k}} q^{\tilde{k} + \frac{\tilde{k}^2}{16}} \leq \frac{(N_u + N_v)^{\tilde{k}}}{\tilde{k}!} \frac{\tilde{k}!}{(N_u + N_v)^{\tilde{k}} n^{\epsilon}} \left(\frac{\tilde{k}!}{(N_u + N_v)^{\tilde{k}} n^{\epsilon}} \right)^{\frac{16}{\tilde{k}^2} \left(\tilde{k} + \frac{\tilde{k}^2}{16} \right) - 1} \leq \frac{1}{n^{\epsilon}}. \tag{26}$$

Finally, combining (25) and (26), we have:

$$\mathbb{E}[\mathrm{Y} \mid \mathrm{H}] \leq \frac{1}{4\tilde{k} n^{\epsilon}} + \frac{1}{n^{\epsilon}} < \frac{2}{n^{\epsilon}}.$$

This finishes the proof of Lemma 18. ◀

A.5 Existences of constants c_2^b and c_3^b

Note that as event H holds, we have $N_u + N_v \leq 6L\rho sn$. Also note that if $\mathsf{K} \leq 12L\rho sn$, then $\tilde{k} \leq 6L\rho sn$. Hence

$$\left(\frac{\tilde{k}!}{\tilde{k}^2 n^{\epsilon} (N_u + N_v)^{\tilde{k}}} \right)^{\frac{1}{\tilde{k}}} > \left(\frac{\sqrt{2\pi\tilde{k}} \tilde{k}^{\tilde{k}} e^{-\tilde{k}}}{\tilde{k}^2 n^{\epsilon}} \right)^{\frac{1}{\tilde{k}}} \frac{1}{6L\rho sn} = \left((2\pi)^{\frac{1}{2\tilde{k}}} \tilde{k}^{-\frac{2}{\tilde{k}}} e^{-1} \right) \left(\frac{1}{n} \right)^{\frac{\epsilon}{\tilde{k}}} \left(\frac{\tilde{k}}{6L\rho sn} \right)$$

$$\left(\frac{\tilde{k}!}{n^{\epsilon} (N_u + N_v)^{\tilde{k}}} \right)^{\frac{16}{\tilde{k}^2}} > (2\pi)^{\frac{8}{\tilde{k}^2}} e^{-\frac{16}{\tilde{k}}} \left(\frac{1}{n} \right)^{\frac{16\epsilon}{\tilde{k}^2}} \left(\frac{\tilde{k}}{6L\rho sn} \right)^{\frac{16}{\tilde{k}}}.$$

Finally, observe that $(2\pi)^{\frac{8}{k^2}} > 1$, $e^{-\frac{16}{k}} > \frac{1}{e}$. It then follows that there exists constants c_2^b, c_3^b such that $c_2^b \left(\frac{1}{n} \right)^{c_3^b/\mathsf{K}} \frac{\mathsf{K}}{sn}$ is smaller than the last term in the right hand side of each equation above. Hence $\tilde{k} \leq 6L\rho sn$ and $q \leq c_2^b \left(\frac{1}{n} \right)^{c_3^b/\mathsf{K}} \frac{\mathsf{K}}{sn}$ implies the condition on q as in Eqn. (22).