Near-Perfect Recovery in the One-Dimensional Latent Space Model

Yu Chen
University of Pennsylvania
chenyu2@cis.upenn.edu

Sampath Kannan
University of Pennsylvania
kannan@cis.upenn.edu

Sanjeev Khanna
University of Pennsylvania
sanjeev@cis.upenn.edu

Abstract

Suppose a graph G is stochastically created by uniformly sampling vertices along a line segment and connecting each pair of vertices with a probability that is a known decreasing function of their distance. We ask if it is possible to reconstruct the actual positions of the vertices in G by only observing the generated unlabeled graph. We study this question for two natural edge probability functions — one where the probability of an edge decays exponentially with the distance and another where this probability decays only linearly. We initiate our study with the weaker goal of recovering only the order in which vertices appear on the line segment. For a segment of length n and a precision parameter δ, we show that for both exponential and linear decay edge probability functions, there is an efficient algorithm that correctly recovers (up to reflection symmetry) the order of all vertices that are at least δ apart, using only $\tilde{O}(\frac{n}{\delta^2})$ samples (vertices). Building on this result, we then show that $O(\frac{n^2 \log n}{\delta^2})$ vertices (samples) are sufficient to additionally recover the location of each vertex on the line to within a precision of δ. We complement this result with an $\Omega(\frac{n^{1.5}}{\delta})$ lower bound on samples needed for reconstructing positions (even by a computationally unbounded algorithm), showing that the task of recovering positions is information-theoretically harder than recovering the order.

2012 ACM Subject Classification Theory of computation → Graph algorithms analysis; Theory of computation → Network optimization

Keywords and phrases One-dimensional latent space model; Reconstruction algorithms; Recovery algorithms.

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

Large graphs arise naturally in modeling many scenarios in social interaction, natural language processing, image processing, and recommendation systems. Nodes in these graphs represent individual entities such as people, genes, or pixels and edges represent relationships between them. A natural goal in analyzing such graphs is to partition the nodes into a small number of sets in such a way that two nodes in the same set 'behave similarly' in terms of their interaction. Algorithms for finding such *communities* are analyzed on synthetic data generated by a stochastic model. The *stochastic block model* or *planted cluster* model is a commonly used generative model. This model is parametrized by (n, k, π, P) where n is the number of vertices, k is the number of clusters, π is a k-vector of probabilities summing to 1, and P is a $k \times k$ matrix. The cluster that a vertex belongs to is chosen independently of other vertices according to π. For any two vertices u and v in clusters i and j respectively, the probability of an edge between u and v is $P[i, j]$. Much work has been done in this model to understand the information-theoretic and computational limits for achieving *exact, partial* and *weak* recovery. For a detailed discussion of the model, its motivation, different notions of recovery, and positive and negative results, see the excellent survey by Abbe [1].

In this paper, we study similar recovery problems in a different model called the *latent space model*. The model was first introduced by Hoff et al. [5] and extended by Handcock et al. [4]. In this model, each node in the graph has a latent position in a Euclidean space, and the relationship of two nodes depends on the distance between them. This model has been applied to political relationships [6, 8] and social networks [3]. Previous work on this model has been focused on algorithmic approaches to finding the maximum likelihood latent positions and empirical evaluations of these approaches [5, 4, 9].

We study the simplest version of the latent space model, where the nodes are uniformly sampled on a segment. We consider both the problem of recovering the order of the nodes and the problem of recovering the positions of the nodes. For this simple setting our focus is on designing algorithms with provable guarantees on number of samples needed, running time, and quality of approximation. Our goal of finding approximate positions for the vertices is also different from the goal of finding the most likely positions.

The stochastic block model is based on the assumption that the entities involved can be neatly categorized into a small number of classes, and membership in a class is the sole determinant of how an entity interacts with others. For example, in this model, we could regard people's political persuasion as being binary – say, liberal or conservative in the United States – and posit that there is a certain probability for edges connecting two conservatives or two liberals, and a different probability for an edge connecting a liberal to a conservative. Many real situations are more complex. For example, the probability of an edge between two nodes in a social network might be a function of many different *attributes* of these nodes, each of which can be discrete or continuous-valued. To model such a generalized view we think of nodes as points in a metric space, and let edges be independently sampled with probabilities that are a decreasing function of the distance between the endpoints. Given a large graph generated according to this model, we seek to find (approximate) locations of each node or entity in the metric space. Our problem formulation can be seen as a generalization of the stochastic block model with equal inter-cluster edge probabilities, by letting the points in the same cluster be at distance 0 from each other, and points in different clusters be at distance 1. In fact, an intermediate model between the stochastic block model and our model consists of a metric space with a finite number of points (or clusters), where each entity is located at one of these points. If we can find good enough approximations for the location of each node in the metric space, we will exactly identify cluster membership in these finite and discrete metric spaces.

In statistical mechanics and probability theory, models such as the one we propose have been studied under the name *long-range percolation models* [11]. Most of the work in these disciplines is focused on the problem of understanding structural properties of the graphs that arise, rather than algorithmic reconstruction of the locations of entities. Our paper takes a first step in designing and analyzing efficient algorithms for this reconstruction. For concreteness and simplicity, we only consider a one-dimensional metric space - the real interval $[0, 1]$. We assume that entities are uniformly sampled (with sufficient density) from this metric space. We also restrict attention to specific types of edge probability functions - exponentially decaying functions and linearly decaying functions. In other words, if d is the distance between points u and v, we consider a model where the probability of an edge is e^{-d} and another model where the probability of an edge is $\frac{1}{d+1}$.

In the stochastic block model, where the problem is to identify the cluster to which each entity belongs, 3 types of recovery are considered: **Exact recovery** where the goal is to identify the cluster membership of every entity with probability close to 1, **Almost exact recovery**, where the goal is to identify the cluster memberships of all but a vanishingly small set of entities with probability close to 1, and **Partial recovery** where the cluster memberships of a constant fraction of the points is determined with probability close to 1. **Weak recovery** is the weakest possible kind of partial recovery, where the fraction of points correctly identified is bounded away from the trivial threshold, which is achieved by an algorithm that ignores the input and randomly guesses the cluster to which each point belongs. In our model, we cannot hope to find the exact location of any point given the finite number of nodes and the fact that locations are only random variables estimated from a stochastic process. Thus, at best we can hope to locate each node only within an interval of some width δ, that depends on the density with which nodes are sampled. With this caveat, we can equivalently define exact, almost exact, partial, and weak recovery. Specifically, in exact (resp. almost exact, partial, weak) recovery, the goal is to approximate the order or the positions of all (resp. almost all, a constant fraction, a non-trivial fraction) of the entities within some constant error.

In the standard stochastic model a distinction is made between fundamental (information-theoretic) limits and (efficient) computational limits for each kind of recovery and bounds for each of them are pretty tightly pinned down. Specifically, the information-theoretic bounds are based on the separation needed between intra-cluster edge probabilities and inter-cluster probabilities. Since our edge probabilities are continuous functions of distance, we cannot hope to show these kinds of bounds. Instead, we give upper and lower bounds for how densely entities must be sampled in order to efficiently recover their approximate order. Since these bounds are essentially tight, and the upper bound is by an efficient algorithm, they are both information-theoretic and computational.

1.1 Problem Statement and Results

We consider the following problem: On the segment $[0, n]$ m points, say $v_1, v_2, \ldots, v_m$, are uniformly sampled. Let x_i be the location of v_i, and let $X = (x_1, x_2, \ldots, x_m)$ be the location vector. A random graph G is constructed with this vertex set; edges are sampled independently as follows: for any pair of vertices v_i and v_j, an edge exists between them with probability $c \cdot f(|x_i - x_j|)$, where c is a number in $(0, 1]$ and f is some monotone decreasing function such that $f(0) = 1$ and $\lim_{x\to\infty} f(x) = 0$. For such a graph G and a position vector X, denote by $\Pr(G|X)$ the likelihood of G given X, i.e. $\Pr(G|X) = \prod_{(i,j)\in G} c \cdot f(|x_i - x_j|) \cdot \prod_{(i,j)\notin G}(1 - c \cdot f(|x_i - x_j|))$.

Our goal is to design an algorithm that takes as input the graph G, and a constant δ, and outputs a vector $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$ which is a “recovery” of the location of each point. We consider two distinct notions of recovery: (1) recovering the order, by which we mean that for any pair of i and j such that $x_i - x_j > \delta$, $\hat{x}_i > \hat{x}_j$ with high probability; (2) recovering the location, by which we mean that for any i, $|x_i - \hat{x}_i| < \delta$ with high probability. We study both these problems for two natural choices of f, namely, the exponential decay function $f(x) = e^{-x}$, and the linear decay function $f(x) = \frac{1}{x+1}$.

For the problem of recovering the order to within any specified precision δ, we show that it suffices to sample $m = \tilde{O}(\frac{n}{\delta^2})$ points. Notice that $\Omega(n \log n)$ points are necessary, since otherwise G will have isolated vertices with high probability, and it is information-theoretically infeasible to determine the relative order of two isolated vertices no matter how far apart.

For the problem of recovering the location, we focus on the case $c = 1$. Building on our algorithm for recovering the order, we can show that with $m = O(n^2 \log n/\delta^2)$ samples, it is possible to recover locations of the points to within precision δ. We also show that the sample complexity of recovering positions is inherently much more than the sample complexity for recovering the order. Specifically, for any $m = o(n^{1.5}/\delta)$, we give two location vectors X^1 and X^2 such that $\|X^1 - X^2\|_\infty > \delta$ and prove that it is impossible to distinguish these two vectors with large constant probability given a random graph G generated in accordance with one of these two vectors. This suggests that $\Omega(n^{1.5}/\delta)$ points are necessary to recover locations. However, given $m = \Omega(n^{1.5} \log n/\delta)$ samples, we prove that we can distinguish between any two location vector X^1 and X^2 such that $\|X^1 - X^2\|_\infty > \delta$. Note that the $\tilde{O}(n^{1.5})$ upper bound refers to the problem of distinguishing two position vectors. The best upper bound we can prove for recovering position is still $\tilde{O}(n^2)$.

Organization: The remainder of the paper is organized as follows. In Section 2, we present and analyze our algorithm for recovering the order of vertices for the exponential decay function. Due to space limitations, we describe our algorithm for the linear decay function in Appendix C. In Section 3, we show that we can recover approximate positions of each vertex in both models. We also establish our lower bound on the number of samples needed for this task. Finally, in Section 4 we briefly discuss the larger context for our problem and open problems.

2 Recovering the Order

We first prove a simple statement — that with enough samples, each segment of length δ has at least one vertex. Throughout the paper, whenever we say $1-o(1)$, we mean $1-1/poly(n)$.

▶ **Lemma 1.** *If $m > \frac{8n\log n}{\delta^2}$ and $\delta < 1$, with probability $1-o(1)$, for any non-negative integer i, the segment $[\frac{i\delta}{2}, \frac{(i+1)\delta}{2}]$ on the segment has at least one point.*

Proof. Since $\log(\frac{1}{\delta}) < \frac{1}{\delta} - 1$, $m > \frac{8n\log n + 8n\log n\log(\frac{1}{\delta})}{\delta} > \frac{8n\log(\frac{n}{\delta})}{\delta}$. For any such segment, the probability that there is no point on it is $(1-\frac{\delta}{2n})^m < e^{-\frac{m\delta}{4n}} = o(\frac{\delta}{n})$. The assertion follows by using the union bound over all segments. ◀

We now give the algorithm that recovers the order for each of the 2 different choices of functions f provided there are sufficiently many vertices. Specifically, we prove the following two theorems. The probability of success indicated in the theorems is over the randomness of the location of the points as well as the realization of the graph.

▶ **Theorem 2.** *When $f(x) = e^{-x}$, for any $0 < \delta < 0.1$ and $m > \frac{2500n\log n}{c^2\delta^2}$, there is a poly-time algorithm that recovers the order with probability $1-o(1)$.*

▶ **Theorem 3.** *When $f(x) = \frac{1}{x+1}$, for any $0 < \delta < 0.1$ and $m > \frac{16000n\log^2 n}{c\delta^2}$, there is a poly-time algorithm that recovers the order with probability $1-o(1)$.*

The basic idea of both algorithms is that, we first approximate the distance between any pair of vertices. The approximation does not need to be very precise in general – we only need the precision when the real distance is within a narrow range. When it is outside that range, the approximation only needs to answer that it is out of range. Since we cannot distinguish between a vector of positions and its reflection, we find a vertex that is very close to an endpoint, and assume that that endpoint is 0, the left end of the segment. Then we use the distance approximations to build the relationship between every pair of vertices that are sufficiently far apart. In other words, for each sufficiently distant pair (u, v), we decide which of u and v is to the left. From these pairwise relationships, we build the order.

We define what we mean by a good approximation of the distance between two vertices.

▶ **Definition 4.** *A distance function $d : V \times V \to \mathbb{R}$ is called a (L, U, δ)-approximation if for any pair of vertices vertices v_i and v_j, $d(v_i, v_j)$ satisfies:*

- *If $|x_j - x_i| < L$, $d(v_i, v_j) < L + \delta$.*
- *If $L \le |x_j - x_i| \le U$, $|x_j - x_i| - \delta < d(v_i, v_j) < |x_j - x_i| + \delta$*
- *If $|x_j - x_i| > U$, $d(v_i, v_j) > U - \delta$.*

We say d is a good approximation if it is an (L, U, δ)-approximation with $3\delta < L < \frac{n}{2} - 2\delta$ and $U > 2L + 8\delta$. We present the algorithm that recovers the order given good approximations. We then present algorithms that produce good approximations for each of the probability functions. (The algorithm for inverse linear decay can be found in Appendix C.)

▶ **Lemma 5.** *There is an algorithm that recovers the order of the vertices if we are given an (L, U, δ)-approximate distance function with $3\delta < L < \frac{n}{2} - 2\delta$ and $U > 2L + 8\delta$ with probability $1 - o(1)$.*

In Section 2.1, we describe such an algorithm. We follow this up with good approximation schemes for $f(x) = e^{-x}$, $f(x) = \frac{1}{x+1}$ in Section 2.2 and Section C respectively.

2.1 Recovering the Order Given Approximation of Distances

In this section, we give an algorithm (ALGORITHM 1) to recover the order of vertices on the segment when we are given a (L, U, δ)-approximate distance function d with $3\delta < L < \frac{n}{2} - 2\delta$ and $U > 2L + 8\delta$. The algorithm works as follows: for any triple of vertices v_i, v_j, and v_k, if v_j is in the middle, then the distance between v_k and v_i is larger than $|x_i - x_j|$ and $|x_j - x_k|$. With a good distance approximation, we can detect which vertex is in the middle, in all triples of vertices that are not too far or too close. We store these ordered triples in a set S (Lemma 6). For any vertex which never occurs in the middle of an ordered triple in S, it must be close to one of the endpoints of the segment. Arbitrarily fixing the position of one such vertex as being near the left endpoint, we can 'recursively orient' each triple in S (Lemma 7), which means that we can tell the order of any vertices that are not too close (Lemma 8). Finally, we use this information to give the full order (Lemma 9). Lemma 5 immediately follows from Lemma 9.

▶ **Lemma 6.** *For any triple (v_i, v_j, v_k) in S, the location of v_j is in the middle of the location of v_i and v_k. On the other hand, for any triple of vertices (v_i, v_j, v_k) such that v_j is in the middle of v_i and v_k, $d(v_i, v_j) \in [L+\delta, 2L+7\delta]$ and $d(v_j, v_k) \in [L+\delta, 2L+7\delta]$, $(v_i, v_j, v_k) \in S$.*

Proof. For any three vertices v_i, v_j, v_k such that $d(v_i, v_j)$ and $d(v_j, v_k)$ both in $[L+\delta, 2L+7\delta]$, we have $|x_i - x_j|$ and $|x_j - x_k|$ are both between L and $2L+8\delta$ by the definition of (L, U, δ) approximation. If v_j is in the middle, then $|x_i - x_k| \ge d(v_i, v_j) + d(v_j, v_k) - 2\delta$, which means $d(v_i, v_k)$ is at least $d(v_i, v_j) + d(v_j, v_k) - 3\delta > |d(v_i, v_j) - d(v_j, v_k)| + 3\delta$ since both of $d(v_i, v_j)$ and $d(v_j, v_k)$ are at least $L > 3\delta$. If v_j is not in the middle, then $|x_i - x_k| \le |d(v_i, v_j) - d(v_j, v_k)| + 2\delta$, which means $d(v_i, v_k) \le |d(v_i, v_j) - d(v_j, v_k)| + 3\delta$. So the triple (v_i, v_j, v_k) is in S if and only if v_j is in the middle. ◀

By Lemma 1 , for any vertex v_j located between $[L + 3\delta, n - L - 3\delta]$, there are two vertices v_i and v_k on its left and its right such that $|x_i - x_j|$ and $|x_j - x_k|$ are both between $L + 2\delta, L + 3\delta$. This means that $d(v_i, v_j)$ and $d(v_j, v_k)$ are both in $[L + \delta, L + 4\delta]$. So $(v_i, v_j, v_k) \in S$ (as $L+4\delta < 2L+7\delta$), which implies vertices in V' are located in $[0, L+3\delta]$ or $[n-L-3\delta, n]$. Furthermore, for any vertex pair (v_i, v_j) with $d(v_i, v_j) \in [L+\delta, 2L+7\delta]$, there

ALGORITHM 1: Order Recovery

For any pair of points v_i and v_j, let $d(v_i, v_j)$ be a (L, U, δ) approximation of $|x_i - x_j|$ with $3\delta < L < \frac{n}{2} - 2\delta$ and $U \geq 2L + 8\delta$;
$S \leftarrow \emptyset$;
for *any triple* (v_i, v_j, v_k) **do**
  if $d(v_i, v_j) \in [L + \delta, 2L + 7\delta] \wedge d(v_j, v_k) \in [L + \delta, 2L + 7\delta] \wedge d(v_i, v_k) > |d(v_i, v_j) - d(v_j, v_k)| + 3\delta$ **then**
    $S \leftarrow S \cup \{(v_i, v_j, v_k)\}$;
$V' \leftarrow \{v \in V | v$ never appears as the middle vertex in any triple in $S\}$;
Pick an arbitrary $v_0 \in V'$;
$V_0 \leftarrow \{v \in V' | d(v_0, v) > U - \delta\}$;
$E' = \{(v_i, v_j) | v_i \in V_0 \wedge d(v_i, v_j) \in [L + \delta, 2L + 7\delta]\}$;
while $S \neq \emptyset$ **do**
  for *any triple* $(v_i, v_j, v_k) \in S$ **do**
    if $(v_i, v_j) \in E'$ **then**
      $E' \leftarrow E' \cup \{(v_j, v_k)\}$;
      $S \leftarrow S - \{(v_i, v_j, v_k), (v_k, v_j, v_i)\}$;
Construct a directed graph $G' = (V, E')$;
For any vertex v, let $R(v)$ be the number of the vertices that can reach v minus the number of vertices reachable from v;
Sort the vertices by $R(v)$ in increasing order and output the order;

exists a vertex v_k such that $(v_i, v_j, v_k) \in S$ or $(v_k, v_j, v_i) \in S$. Without loss of generality, suppose $v_0 \in [n - L - 3\delta, n]$. Then V_0 contains all the vertices v_j such that no vertex v_i on its left with $d(v_i, v_j) \in [L + \delta, 2L + 7\delta]$.

▶ **Lemma 7.** *The while loop of the algorithm always terminates. Moreover, for any pair of vertices v_i and v_j, $(v_i, v_j) \in E'$ if and only if v_i is to the left and $d(v_i, v_j) \in [L + \delta, 2L + 7\delta]$.*

Proof. We first prove that for any pair of vertices (v_i, v_j) in E', v_i is to the left of v_j, using induction on the order of the pairs added to E'. For the base case, V_0 only contains vertices with no vertex on their left with approximate distance at least $L + \delta$. So for any pair (v_i, v_j) added into E' before the while loop, v_i is to the left. Assume inductively that this is true for all pairs added before the current iteration of the while loop. For any pair (v_i, v_j) added into E' in the current iteration, there is a vertex v_i' such that $(v_i', v_i, v_j) \in S$ and $(v_i', v_i) \in E'$. By induction hypothesis, v_i' is on v_i's left. So v_i is between v_i' and v_j, so v_i is on v_j's left by Lemma 6.

We prove that the while loop terminates, i.e., that all triples in S eventually get deleted. Suppose for contradiction that, v_i is the leftmost vertex to appear in any undeleted triple, and there is a triple (v_i, v_j, v_k) that never gets deleted. (Note that whenever $(v_k, v_j, v_i) \in S$, $(v_i, v_j, v_k) \in S$). If there exists a vertex v_i' to the left of v_i with $d(v_i', v_i) \in [L+\delta, 2L+7\delta]$, then (v_i', v_i, v_j) is in S and will be deleted sometime, then $(v_i, v_j) \in E'$, which means (v_i, v_j, v_k) will be deleted. If there is no such vertex v_i' then $v_i \in V_0$, which also means $(v_i, v_j) \in E'$, (v_i, v_j, v_k) will be deleted in the first iteration. Thus contradicts that (v_i, v_j, v_k) would never gets deleted.

Finally, we prove that any pair of vertices (v_i, v_j) with $d(v_i, v_j) \in [L + \delta, 2L + 7\delta]$ will be added into E'. This is because by Lemma 1 , there exists a vertex v_k such that $(v_i, v_j, v_k) \in S$

or $(v_k, v_j, v_i) \in S$. Since such triple was deleted in the while loop, (v_i, v_j) has been added into E'. ◀

▶ **Lemma 8.** *For any pair of vertices v_i and v_j, the vertex v_j is reachable from v_i in G' if and only if $d(v_i, v_j) \geq L + \delta$ and v_i is to the left.*

Proof. If v_j is reachable from v_i, there is a path form v_i to v_j, and the location of any vertex on the path is to the left of the next vertex on the path. So v_i is on v_j's left. If $(v_i, v_j) \in E'$, by Lemma 7, $d(v_i, v_j) \geq L + \delta$, otherwise the path has at least three vertices. By Lemma 7, any neighbouring vertex has distance at least L, which means the distance between v_i and v_j is at least $2L$, so $d(v_i, v_j) \geq 2L - \delta > L + \delta$.

For any pair v_i, v_j with v_i to the left and $d(v_i, v_j) \geq L + \delta$, if $d(v_i, v_j) \leq 2L + 7\delta$, then $(u_i, v_j) \in E'$, which means v_j is reachable from v_i in G'. If $d(v_i, v_j) > 2L + 7\delta$, then the distance between them is at least $2L + 6\delta$. by Lemma 1, there exists a sequece of vertex $v_i = u_1, u_2, \ldots, u_k = v_j$ such that for any $1 \leq \ell \leq k - 1$, u_ℓ is to the left of $u_{\ell+1}$, and the distance between them is between $L + 2\delta$ and $2L + 6\delta$, which means $d(u_\ell, u_{\ell+1}) \in [L+\delta, 2L+7\delta]$, in other words, by Lemma 7, $(u_\ell, u_{\ell+1}) \in E'$, so v_j is reachable form v_i in G'. ◀

▶ **Lemma 9.** *The output order of the algorithm satisfies that for any v_i and v_j that are separated by a distance of at least 3δ, v_i appears prior to v_j in the order if and only if v_i is to the left of v_j.*

Proof. If v_i is to the left and the distance between v_i and v_j is at least 3δ, for any vertex v_k on v_j's right with $d(v_j, v_k) \geq L + \delta$, we have $x_k - x_j \geq L$, which means $x_k - x_i \geq L + 3\delta$ and $d(v_i, v_k) \geq L + 2\delta$. For any vertex v_k on v_i's left with $d(v_i, v_k) \geq L + \delta$, $x_i - x_k \geq L$, which means $x_j - x_k \geq L + 3\delta$ and $d(x_k, x_j) \geq L + 2\delta$. So $R(x_i) \leq R(v_j)$. On the other hand, by Lemma 1 and the fact that $L < \frac{n}{2} - 2\delta$, there exists a vertex v_k with one of the following two properties:

- v_k is on v_j's right and $x_k - x_j < L$ and $x_k - x_i > L + 2\delta$.
- v_k is on v_i's left and $v_i - v_k < L$ and $v_j - v_k > L + 2\delta$.

In the first case, $d(v_j, v_k) < L + \delta$ while $d(v_i, v_k) > L + \delta$, which means v_k is reachable from v_i but not v_j. In the second case, $d(v_i, v_k) < L+\delta$ while $d(v_j, v_k) > L+\delta$, which means v_j is reachable from v_k but v_i is not reachable from v_k. So $R(v_j)$ is strictly larger than $R(v_i)$. ◀

2.2 Distance Approximation for Exponential Decay Function

In this section, we consider the case that $f(x) = e^{-x}$. The probability of an edge between two vertices v_i and v_j, with locations x_i and x_j respectively, is $c \cdot e^{-|x_i - x_j|}$. We first analyze the degree of each vertex and the number of common neighbors between each pair of vertices.

▶ **Lemma 10.** *For any vertex v_i located at position x_i on the segment, if we uniformly sample a vertex v on the segment, then the edge (v_i, v) is present with probability $\frac{c}{n}(2 - e^{-x_i} - e^{x_i - n})$. In other words, this is the expected probability of an edge from v_i, where the expectation is over the choice of the other endpoint v.*

▶ **Lemma 11.** *For any two vertices v_i and v_j located at x_i and x_j respectively with $x_i < x_j$, if we uniformly sample a vertex v on the segment, then v is a common neighbor of v_i and v_j with probability $\frac{c^2}{n}((x_j - x_i + 1)e^{x_i - x_j} - \frac{1}{2}(e^{x_i + x_j - 2n} + e^{-x_i - x_j}))$.*

By Lemma 11, the number of common neighbors of a pair of vertices “mostly” depends on the distance between these two vertices. We use the degree of these two vertices to eliminate the effect of the remaining terms. We first prove that we can check if two vertices are far away.

▶ **Lemma 12.** *If $m > \frac{2500n\log n}{c^2\delta^2}$, with probability $1-o(1)$, for any two vertices v_i and v_j, (a) if they have no common neighbor, then $|x_i - x_j| > 2.5$, and (b) if $|x_i - x_j| > n/2$, then they have no common neighbor.*

We now describe how to approximate the distance between two vertices.

▶ **Lemma 13.** *If $0 < \delta < 0.1$ and $m > \frac{2500n\log n}{c^2\delta^2}$, then for any pair of vertices v_i and v_j, with probability $1 - O(n^{-2.5})$, we can calculate $\hat{d}$, an approximation of $d = |x_i - x_j|$ such that:*

- *If $d < 0.3$, $\hat{d} < 0.3 + \delta$.*
- *If $0.3 \le d \le 2.5$, $d - \delta < \hat{d} < d + \delta$*
- *If $d > 2.5$, $\hat{d} > 2.5 - \delta$.*

Proof. For any number x, let $g(x) = (x+1)e^{-x}$ and $h(x) = e^{-x} + e^{x-n}$. We first prove that we can either approximate $g(d)$ with additive error at most $0.2d$ or directly output a $\hat{d}$ which satisfies the condition.

We first check if v_i and v_j have common neighbors. If they have no common neighbor, then by Lemma 12, $d > 2.5$. So we can directly output $\hat{d} = n$. Otherwise we have $d < n/2$.

By Lemma 11 and Proposition 34, we can approximate $g(d) + \frac{1}{2}(e^{x_i+x_j-2n} + e^{-x_i-x_j})$ with additive error $\frac{\delta}{11}$ since $m > \frac{2500n\log n}{c^2\delta^2}$. To eliminate the terms $e^{x_i+x_j-2n}$ and $e^{-x_i-x_j}$, we use the degree of v_i and v_j. By Lemma 10 and Proposition 34, we can approximate $h(x_i)$ and $h(x_j)$ with additive error $\frac{\delta}{11}$. On the other hand, $h(x_i)\cdot h(x_j) = e^{-x_i-x_j} + e^{x_i+x_j-2n} + e^{-n+x_i-x_j} + e^{-n-x_i+x_j}$. The last two terms are $o(1)$ since $|x_i - x_j| < n/2$. So we can approximate $e^{-x_i-x_j} + e^{x_i+x_j-2n}$ with additive error $\frac{2\delta}{11} + o(1) < \frac{\delta}{5}$. We can thus approximate $g(d)$ with additive error at most $\frac{\delta}{5}$.

The proof is completed by the observation that $g(x)$ is monotone decreasing when $x \ge 0$, and the derivative $g'(x) < -0.2$ when $0.3 \le x \le 2.5$. ◀

Note that if $0 < \delta < 0.1$, $3\delta < 0.3 < \frac{n}{2} - 2\delta$ and $2.5 > 0.3*2+8\delta$. Theorem 2 immediately follows from Lemma 5 and Lemma 13.

3 Recovering the Position

In this section, we consider the problem of recovering the positions of the vertices on the segment. First, we prove the following simple result, which extends the results for recovering the order.

▶ **Lemma 14.** *Suppose $m > \frac{10n^2\log n}{\delta^2}$. For any function f, if we can recover the order of the vertices, then we can also recover a position vector $\hat{X}$ such that for any i, $|x_i - \hat{x}_i| < 2\delta$ with probability $1-o(1)$.*

Proof. Suppose the order output by the order recovery algorithm is $(v_1, v_2, \ldots, v_m)$, and their true positions are $(x_1, x_2, \ldots, x_m)$. We will prove that $\left|x_i - \frac{in}{m}\right| < 2\delta$ (i.e. we can just output the position as uniformly dispersed along the segment according to the order).

Suppose the real order is $(u_1, u_2, \ldots, u_m)$, and the real positions are $(y_1 < y_2 < \cdots < y_m)$. We first prove $|x_i - y_i| < \delta$, and then prove that $\left|y_i - \frac{in}{m}\right| < \delta$. The following arguments are based on the event that the run of the order recovery algorithm is successful.

For any i, if $x_i - y_i \geq \delta$, then for any $j \leq i$, $x_i - y_j \geq \delta$. By the definition of recovering the order, for any $j \leq i$, u_j occurs before v_i in the order output by the algorithm, which contradicts the fact that v_i appears at the i^{th} position of the order output by the algorithm. So $x_i - y_i < \delta$. For the same reason, we also have $y_i - x_i < \delta$.

On the other hand, for any $1 \leq k \leq \frac{2n}{\delta}$, let Z_k be the number of vertices sampled in segment $[0, k\delta/2]$. By the Chernoff bound, with probability $1 - o(\frac{1}{n})$, $\left|Z_k - \frac{km\delta}{2n}\right| < \frac{m}{2\delta n}$. By taking the union bound over the complementary events, all Z_k's are close to their expectation with probability $1 - o(1)$. For any i, suppose $\frac{(k-1)m\delta}{2n} < i \leq \frac{km\delta}{2n}$, then there are at most i vertices sampled in the segment $[0, (k-2)\delta/2]$ and at least i vertices sampled in the segment $[0, (k+1)\delta/2]$, which implies $(k-2)\delta/2 < y_i < (k+1)\delta/2$. On the other hand, $(k-1)\delta/2 < i \leq k\delta/2$, so $\left|y_i - \frac{in}{m}\right| < \delta$. ◀

By Lemma 14 and the results in Section 2, we can recover the position with $\tilde{\Omega}(n^2)$ vertices for both choices of f. However, there is a huge gap compared to the number of samples necessary for recovering the order.

In the remainder of this section, we consider the following "weaker" problem: the task is distinguishing two position vectors $X = (x_1, x_2, \ldots, x_m)$ and $Y = (y_1, y_2, \ldots, y_m)$ with the guarantee that vertices in X and Y have the same order. We focus on the exponential decay function $f(x) = e^{-x}$ and the case when the number of samples is between the gap of Theorem 2 and Lemma 14. Say two position vectors X and Y are δ-far if there exists a vertex v_i such that $|x_i - y_i| > \delta$. We prove that we cannot distinguish two positions which are δ far away when there are $o(n^{1.5})$ samples. This shows that we cannot recover the position of vertices with only $o(n^{1.5})$ samples even if the algorithm is given the order.

▶ **Theorem 15.** *For any $1000n < m < \frac{(10^{-5})n^{1.5}}{\delta}$, if X is sampled uniformly at random, then with probability $1-o(1)$, we can construct a position vector Y which has the same order as X and δ-far from X such that, if we randomly sample a graph G according to X, there is a constant probability that $\Pr(G|X) < \Pr(G|Y)$.*

On the other hand, we prove that if $m = \Omega(n^{1.5}\log n)$, then we can distinguish any two position vectors which are far from each other when one vector is sampled uniformly, which means Theorem 15 is tight up to a $O(\log n)$ factor.

▶ **Theorem 16.** *For any $\frac{n^{1.5}\log n}{\delta} < m = \tilde{O}(n^2)$, if X is sampled uniformly at random, then with probability $1-o(1)$, for any position vector Y with the same vertex order and δ-far from X, suppose we randomly sample a graph G according to X, then with probability $1-o(1)$, $\Pr(G|X) > \Pr(G|Y)$.*

We prove Theorem 15 in Section 3.1, and prove Theorem 16 in Section 3.2.

3.1 Proof of Theorem 15

For any graph G and two position vectors X and Y, $\Pr(G|X) > \Pr(G|Y)$ if and only if $\log\Pr(G|X) > \log\Pr(G|Y)$, which means $\sum_{(v_i,v_j)\in G}\log(e^{-|x_i-x_j|}) + \sum_{(v_i,v_j)\notin G}\log(1 - e^{-|x_i-x_j|})$ is larger than $\sum_{(v_i,v_j)\in G}\log(e^{-|Y_i-Y_j|}) + \sum_{(v_i,v_j)\notin G}\log(1 - e^{-|y_i-y_j|})$.

Let $L = \log\Pr(G|X) - \log\Pr(G|Y)$ and $L_{i,j} = \log(e^{-|x_i-x_j|}) - \log(e^{-|y_i-y_j|})$ if $(v_i, v_j) \in G$ and $L_{i,j} = \log(1 - e^{-|x_i-x_j|}) - \log(1 - e^{-|y_i-y_j|})$ if $(v_i, v_j) \notin G$. The probability that $\Pr(G|X) > \Pr(G|Y)$ is equal to the probability that $L = \sum_{i,j} L_{i,j} > 0$.

Without loss of generality, suppose $x_1 < x_2 < \cdots < x_m$. Let Y be the position vector $(y_1, y_2, \ldots, y_m)$ such that $y_i = (1 - \frac{2\delta}{n})x_i$. It is easy to see that as long as m is super constant, $|x_m - y_m| > \delta$ with probability $1 - o(1)$, which means X and Y are δ-far.

The proof of Theorem 15 has the following steps. We first prove that the expectation of L is small (roughly speaking, we prove that it is much smaller than its deviation). Thus by anti-concentration bound (Proposition 38), with some constant probability, $|L - \mathbb{E}[L]| > \mathbb{E}[L]$. Then we prove that L is also not so far from $\mathbb{E}[L]$ by a concentration bound (Proposition 36), which guarantees that the probability of $L - \mathbb{E}[L] > \mathbb{E}[L]$ and $\mathbb{E}[L] - L > \mathbb{E}[L]$ are roughly equal. This means that there is a constant probability that $L < 0$.

However, if v_i and v_j are far away in X, $L_{i,j}$ has a very large deviation. Thus we cannot use a concentration bound to bound the sum of these $L_{i,j}$. To solve this problem, we first prove that the sum of $L_{i,j}$ where v_i and v_j are not too far (we denote the sum as $\bar{L}$) is close to L, and then analyze $\bar{L}$ instead of L.

Throughout this section, we let $d_{i,j} = |x_i - x_j|$ and $d'_{i,j}$ as $|x_i - x_j| - |y_i - y_j|$.

We first bound $L_{i,j}$ when v_i and v_j are very far away.

▶ **Lemma 17.** *If $|x_i - x_j| \geq n^{0.1}$, then $|L_{i,j}| < n^{-100}$.*

Denote $i \sim j$ if $|x_i - x_j| < n^{0.1}$ and $\bar{L} = \sum_{i \sim j} L_{i,j}$. By Lemma 17, $L - \bar{L} < n^{-90}$ by taking union bound on all pairs of i and j. So in order to prove that $\Pr(L < 0) = \Omega(1)$, it is sufficient to prove $\Pr\left(\bar{L} < -n^{-90}\right) = \Omega(1)$.

The following lemma gives some properties of $\mathbb{E}[L_{i,j}]$.

▶ **Lemma 18.** *For any pair of vertices v_i, v_j, $e^{-d_{i,j}}(d'^2_{i,j}/2 + (1 - e^{-d_{i,j}})a/2) < \mathbb{E}[L_{i,j}] < e^{-d_{i,j}}(d'^2_{i,j} + \frac{2d'^2_{i,j}}{d_{i,j}})$ where $a = \frac{e^{-d_{i,j}}(e^{d'_{i,j}} - 1)}{1 - e^{-d_{i,j}}}$.*

Next, we give an upper bound for $\mathbb{E}[L]$.

▶ **Lemma 19.** *If $m < \frac{(10^{-5})n^{3/2}}{\delta}$ and X is obtained by sampling each point uniformly, then $\mathbb{E}\left[\sum_{i,j} L_{i,j}\right] < 10^{-8}$ with probability $1 - o(1)$.*

By Lemma 18, $\mathbb{E}[L_{i,j}]$ is always positive, so $\mathbb{E}\left[\bar{L}\right] < \mathbb{E}[L] < 10^{-8}$. Then we use Proposition 38 to prove there is a constant probability that $\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| = \Omega(\sqrt{\mathbb{E}\left[\bar{L}\right]})$.

▶ **Lemma 20.** $\Pr\left(\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| > 10^{-3}\sqrt{\mathbb{E}\left[\bar{L}\right]}\right) \geq 0.5.$

We now prove that $\bar{L}$ is not very far from its expectation. We first prove that if v_i and v_j are not far apart, then $L_{i,j}$ is sub-exponential random variable (see Definition 35).

▶ **Lemma 21.** *For any pair of i and j, if $d_{i,j} < n^{0.1}$, then $L_{i,j}$ is a sub-exponential variable with parameters $(\sigma_{i,j}, b)$ where $\sigma^2_{i,j} = 48\,\mathbb{E}[L_{i,j}]$ and $b = n^{-0.8}$.*

We also need a very loose lower bound on $\mathbb{E}\left[\bar{L}\right]$.

▶ **Lemma 22.** *If $1000n < m < \frac{(10^{-5})n^{3/2}}{\delta}$, then $\mathbb{E}\left[\bar{L}\right] = \omega(n^{-1.6})$.*

We are ready to use Proposition 36 to prove the concentration of $\bar{L}$.

▶ **Lemma 23.** *If $1000n < m < \frac{(10^{-5})n^{3/2}}{\delta}$, then for any integer $k > 0$, $\Pr\left(\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| > 10k\sqrt{\mathbb{E}[\bar{L}]}\right) < 4e^{-k}$.*

Finally, we put all the results together,

Proof of Theorem 15. By Lemma 17, it is sufficient to prove $\Pr\left(\bar{L} < -n^{-90}\right) = \Omega(1)$. By Lemma 20, $\Pr\left(-n^{-90} < \bar{L} < (10^{-3})\sqrt{\mathbb{E}\left[\bar{L}\right]} - n^{-90}\right) < 0.5$. So

$$\int_{-n^{-90}}^{((10)^{-3})\sqrt{\mathbb{E}[\bar{L}]}-n^{-90}} x\Pr\left(\bar{L} = x\right) dx > -0.5n^{-90}$$

By Lemma 23, $\Pr\left(\bar{L} < -(10k)\sqrt{\mathbb{E}\left[\bar{L}\right]}\right) < 4e^{-k}$ for any integer $k > 0$, which means

$$\begin{aligned}\int_{-\infty}^{-200\sqrt{\mathbb{E}[\bar{L}]}} x\Pr\left(\bar{L} = x\right) dx &> \sum_{k=20}^{\infty} -\frac{(10k+1)\sqrt{\mathbb{E}\left[\bar{L}\right]}}{e^k} \\ &= -(10e^{-20})(\frac{1}{(1-e^{-1})^2} + \frac{1}{1-e^{-1}}) \cdot \sqrt{\mathbb{E}\left[\bar{L}\right]} \\ &> -(10^{-7})\sqrt{\mathbb{E}\left[\bar{L}\right]}\end{aligned}$$

Let $P_1 = \Pr\left(\bar{L} > (10^{-3})\sqrt{\mathbb{E}\left[\bar{L}\right]} - n^{-90}\right)$, then

$$\int_{(10^{-3})\sqrt{\mathbb{E}[\bar{L}]}-n^{-90}}^{\infty} x\Pr\left(\bar{L} = x\right) dx \geq P_1(10^{-3})\sqrt{\mathbb{E}\left[\bar{L}\right]} - P_1 n^{-90}$$

Moreover, let $\Pr\left(-200\sqrt{\mathbb{E}\left[\bar{L}\right]} \leq \bar{L} < n^{-90}\right) = P_2$, then

$$\int_{-200\sqrt{\mathbb{E}[\bar{L}]}}^{n^{-90}} x\Pr\left(\bar{L} = x\right) dx > -200P_2\sqrt{\mathbb{E}\left[\bar{L}\right]}$$

By Lemma 19,

$$\begin{aligned}(10^{-4})\sqrt{\mathbb{E}\left[\bar{L}\right]} < \mathbb{E}\left[\bar{L}\right] &= \int_{-\infty}^{\infty} x\Pr\left(\bar{L}\right) dx \\ &< ((10^{-3})P_1 - (10^{-7}) - 200P_2)\sqrt{\mathbb{E}\left[\bar{L}\right]} - (0.5 + P_1)n^{-90}\end{aligned}$$

So $10P_1 - 10^{-3} - 2000000P_2 - o(1) < 1$ by Lemma 22, which implies $10P_1 - 2000000P_2 < 1.1$. On the other hand, since $\Pr\left(\bar{L} < 199\sqrt{\mathbb{E}\left[\bar{L}\right]}\right) < e^{-20}$, so $P_1 + P_2 > 1 - 0.5 - e^{-20} > 0.4$. So $P_2 = \Omega(1)$. ◀

3.2 Proof of Theorem 16

We define $L_{i,j}$ and L as in Section 3.2. To prove Theorem 16, we need to prove $\Pr(L > 0) = 1 - o(1)$. The basic idea is to prove $\mathbb{E}[L]$ is large and use the concentration bound (Propostion 36) to prove $\mathbb{E}[L]$ is larger than the "concentration range".

Although we also prove the concentration of L in Section 3.2, the difference is that, here the second position vector Y is selected by an adversary. Some $L_{i,j}$'s might be "ill-behaved" and thus their deviation is hard to control due to the choice of Y. To solve this problem, we construct $\bar{L}_{i,j}$ as follows: If $|y_i - y_j| > |x_i - x_j|$, then let $\bar{L}_{i,j} = \min\{2, L_{i,j}\}$ if $(v_i, v_j) \in G$; if $|y_i - y_j| < |x_i - x_j|$, then let $\bar{L}_{i,j} = (1 - e^{-L_{i,j}}) + \frac{1}{2}(1 - e^{-L_{i,j}})^2$; if $(v_i, v_j) \notin G$.

In any scenerio, $\bar{L}_{i,j}$ is always smaller than $L_{i,j}$. (This is due to Proposition 29.) So $\Pr\left(\sum_{i,j} \bar{L}_{i,j} > 0\right) \le \Pr(L > 0)$. Moreover, let $\bar{L}$ be the sum of $\bar{L}_{i,j}$ excluding those pairs i and j where $|x_i - x_j| > 5\log n$ and $|x_i - x_j| > |y_i - y_j|$. For such pairs (i,j), the probability that $(v_i, v_j) \notin G$ is $1 - O(n^{-5})$ and in that event, $\bar{L}_{i,j} > 0$. Since there are at most $m^2 = o(n^5)$ pairs of such (i,j), with probability $1 - o(1)$, all of these $\bar{L}_{i,j}$'s are greater than 0. So with probability $1 - o(1)$, $\bar{L} \le \sum_{i,j} \bar{L}_{i,j} \le L$. So it is sufficient to prove $\Pr\left(\bar{L} > 0\right) = 1 - o(1)$. We call the unexcluded pairs as the pair contributing to $\bar{L}$.

Throughout this section, let $d_{i,j} = |x_i - x_j|$ and $d'_{i,j}$ as $||x_i - x_j| - |y_i - y_j||$.

We first prove a simple lemma about the distance between each pair of vertices in X.

▶ **Lemma 24.** *If $m = \tilde{O}(n^2)$, with probability $1 - o(1)$, for any pair of (i,j), $|x_i - x_j| > \frac{1}{n^4}$.*

Hereafter, we assume $d_{i,j} > \frac{1}{n^4}$ for all pair of i and j. We prove the following property of $\bar{L}_{i,j}$.

▶ **Lemma 25.** *For pairs (i,j) that contribute to $\bar{L}$, $\bar{L}_{i,j}$ is a sub-exponential random variable with parameter $(\sigma_{i,j}, b)$ where $\sigma^2_{i,j} = 10\log n \cdot \mathbb{E}\left[\bar{L}_{i,j}\right]$ and $b = 10\log n$.*

Next, we analyze the expectation of $\bar{L}$. The following lemma is a byproduct of the proof of Lemma 25.

▶ **Lemma 26.** *For any i, j, $\mathbb{E}\left[\bar{L}_{i,j}\right] > \frac{1}{6}e^{-d_{i,j}}d'^2_{i,j}$ if $d'_{i,j} \le 2$. Otherwise $\mathbb{E}\left[\bar{L}_{i,j}\right] > e^{-d_{i,j}}$.*

We prove a lower bound on the expectation of $\bar{L}$.

▶ **Lemma 27.** *If $\frac{100n^{1.5}\log n}{\delta} < m = \Omega(n^2)$. If X is sampled uniformly, then with probability $1 - o(1)$, for any Y such that there is a pair of i and j satisfies that $d'_{i,j} > \frac{\delta}{2}$, $\mathbb{E}\left[\bar{L}\right] > 5\log^2 n$.*

Now we are ready to use the concentration bound (Proposition 36) to prove Theorem 16.

Proof of Theorem 16. Let v_j (resp. v_k) be the left (resp. right) most vertex in X, then with probability $1 - o(1)$ $x_j = o(1)$ and $x_k = n - o(1)$. Let v_i be the vertex such that $|x_i - y_i| > \delta$, then either $d'_{i,j} > \delta - o(1)$ or $d'_{i,k} > \delta - o(1)$. Suppose $d'_{i,j} > \delta - o(1) > \frac{\delta}{2}$. By Lemma 27, $\mathbb{E}\left[\bar{L}\right] > 5\log^2 n$. By Lemma 25 and Proposition 36,

$$
\begin{aligned}
\Pr\left(\bar{L} < 0\right) &\le \Pr\left(\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| > \mathbb{E}\left[\bar{L}\right]\right) \\
&< 2e^{-\frac{\mathbb{E}[\bar{L}]^2}{20\,\mathbb{E}[\bar{L}]\log n}} = 2e^{-\frac{\mathbb{E}[\bar{L}]}{20\log n}} < 2e^{-1.25\log n} \\
&= o(1)
\end{aligned}
$$

◀

4 Conclusions

We developed a framework for recovery that uses the following high-level approach: 1) use the graph to reconstruct approximate degrees and common neighborhood sizes for pairs of vertices; 2) use this information to approximately identify the neighborhoods of each vertex, and spatial relationships between vertices in each neighborhood; and finally, 3) use the local knowledge to establish global structure - order relations or positions. Using this framework, we obtained essentially tight bounds on the number of samples required for recovering the (approximate) order of points on a line segment under both exponential decay and linear decay models. It would be interesting to close the gap that remains between the upper and lower bounds for recovering the location of the points.

This paper can be seen as taking the first step in what should be a promising line of research, that will include generalizing our results to other metric spaces as well as to other edge probability functions. As we move from one-dimensional space to higher dimensional spaces, recovery becomes distinctly harder (as one might expect) but our preliminary investigation suggests that the framework described in this work continues to be of value in understanding recovery in $\mathbb{R}^k$ for $k \geq 2$. Beyond this, a particularly intriguing problem is to recover missing attributes. If we are given a graph as well as some partial information about the attributes of vertices, can we learn both the edge probability function and values of the missing attributes? Such problems are likely to be of interest in social science research, as well as in understanding diverse networks such as biological and economic networks.

References

1. E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.
2. S. Bernstein. On a modification of chebyshevs inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.
3. J. C. Fisher. Social space diffusion: Applications of a latent space model to diffusion with uncertain ties. *Sociological Methodology*, page 0081175018820075, 2019.
4. M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.
5. P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.
6. P. D. Hoff and M. D. Ward. Modeling dependencies in international relations networks. *Political Analysis*, 12(2):160–175, 2004.
7. A. Kolmogorov. Sur les propriétés des fonctions de concentrations de mp lévy. *Ann. Inst. H. Poincaré*, 16(1):27–34, 1958.
8. T. L. J. Ng, T. B. Murphy, T. Westling, T. H. McCormick, and B. K. Fosdick. Modeling the social media relationships of irish politicians using a generalized latent space stochastic blockmodel. *arXiv preprint arXiv:1807.06063*, 2018.
9. A. E. Raftery, X. Niu, P. D. Hoff, and K. Y. Yeung. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919, 2012.
10. B. Rogozin. An estimate for concentration functions. *Theory of Probability & Its Applications*, 6(1):94–97, 1961.
11. L. S. Schulman. Long range percolation in one dimension. *Journal of Physics A: Mathematical and General*, 16(17):L639, 1983.
12. A. Zygmund et al. Hg hardy, je littlewood and g. pólya, inequalities. *Bulletin of the American Mathematical Society*, 59(4):411–412, 1953.

A Math Tools

A.1 Basic Math Inequalities

In this section, we prove some math results which we will use.

▶ **Proposition 28.** *Suppose four different numbers a, a', b, b', ε satisfy that $0 \leq \varepsilon < 1/2$, $|a - a'| < \varepsilon a$, $|b - b'| < \varepsilon b$, and $4 < a < b$, then* $\left|\frac{\log b - \log a}{b-a} - \frac{\log b' - \log a'}{b'-a'}\right| < \varepsilon$

Proof. For any positive numbers i, j, let $g(i,j) = \frac{\log i - \log j}{j-i}$. Then $g(i,j) = \int_i^j \frac{1}{x} dx$, which means $g(i,j)$ is between $\frac{1}{i}$ and $\frac{1}{j}$.

We first prove $|g(a,b) - g(a',b)| < \frac{\varepsilon}{2}$, and with the same argument, $|g(a',b) - g(a',b')| < \frac{\varepsilon}{2}$, which together imply the proposition.

Case 1: $a' < a < b$. $g(a',b) = \frac{b-a}{b-a'}g(a,b) + \frac{a-a'}{b-a'}g(a,a')$, which means $|g(a',b) - g(a,b)| = \frac{a-a'}{b-a'}\,|g(a,a') - g(a,b)| < \frac{a-a'}{b-a'}(\frac{1}{a'} - \frac{1}{b}) = \frac{a-a'}{a'b} < \frac{2\varepsilon}{b} < \frac{\varepsilon}{2}$.

Case 2: $a < a' < b$. $g(a,b) = \frac{b-a'}{b-a}g(a',b) + \frac{a'-a}{b-a}g(a',a)$, which means $|g(a',b) - g(a,b)| = \frac{a-a'}{b-a}\,|g(a,a') - g(a',b)| < \frac{a'-a}{b-a}(\frac{1}{a} - \frac{1}{b} = \frac{a'-a}{ab} < \frac{\varepsilon}{b} < \frac{\varepsilon}{4}$.

Case 3: $a < b < a'$, $|g(a,a') - g(b,a')| < \frac{1}{a} - \frac{1}{a'} < \frac{\varepsilon}{a'} < \frac{\varepsilon}{4}$. ◀

▶ **Proposition 29.** *If* $0 < x$, $x + x^2/2 < \log(1-x)$*; if* $x < 0.5$, $\log(1-x) < x + x^2$.

Proof. The Taylor expansion of $\log(1-x)$ is

$$-\log(1-x) = \sum_{k=1}^{\infty} \frac{x^k}{k} > x + x^2/2$$

The inequality is because $x > 0$. On the other hand,

$$\sum_{k=1}^{\infty} \frac{x^k}{k} < x + \frac{1}{2}\sum_{k=2}^{\infty} x^k < x + x^2$$

since $x < 0.5$. ◀

▶ **Proposition 30.** *For any* $0 < x' \le x$, $\frac{e^{-x}(e^{x'}-1)}{1-e^{-x}} \le \frac{x'}{x}$.

Proof. Let $\varepsilon = \frac{x'}{x}$, to prove the proposition, we only need to prove that for any $0 < \varepsilon \le 1$, $\frac{e^{-x}(e^{\varepsilon x}-1)}{1-e^{-x}} < \varepsilon$, which is equivalent to prove that

$$e^{(\varepsilon-1)x} - (1-\varepsilon)e^{-x} < \varepsilon$$

Let $f_\varepsilon(x)$ be the LHS, $f_\varepsilon(0) = \varepsilon$, and the derivative

$$f'_\varepsilon(x) = (\varepsilon-1)e^{(\varepsilon-1)x} - (\varepsilon-1)e^{-x} < 0$$

when $x > 0$, so $f_\varepsilon(x) < \varepsilon$ when $x > 0$. ◀

▶ **Proposition 31.** *For any* $0 < x' \le x$, $\frac{e^{-x}(1-e^{-x'})}{1-e^{-x}} \le \frac{x'}{x}$.

Proof. Let $\varepsilon = \frac{x'}{x}$, to prove the proposition, we only need to prove that for any $\varepsilon > 0$, $\frac{e^{-x}(1-e^{-\varepsilon x})}{1-e^{-x}} < \varepsilon$, which is equivalent to prove that

$$(1+\varepsilon)e^{-x} - e^{-(\varepsilon+1)x} < \varepsilon$$

Let $f_\varepsilon(x)$ be the LHS, $f_\varepsilon(0) = \varepsilon$, and the derivative

$$f'_\varepsilon(x) = -(\varepsilon+1)e^{-x} + (\varepsilon+1)e^{-(\varepsilon+1)x} < 0$$

when $x > 0$, so $f_\varepsilon(x) < \varepsilon$ when $x > 0$. ◀

▶ **Proposition 32.** *For any* $x' > x$, $\frac{1-e^{-x'}}{1-e^{-x}} < \frac{x'}{x}$.

Proof. Let $\varepsilon = \frac{x'}{x}$, to prove the proposition, we only need to prove that for any $\varepsilon > 1$, $\frac{1-e^{-\varepsilon x}}{1-e^{-x}} < \varepsilon$, which is equivalent to prove that

$$e^{-\varepsilon x} - \varepsilon e^{-x} + \varepsilon - 1 > 0$$

Let $f_\varepsilon(x)$ be the LHS, $f_\varepsilon(0) = 0$, and the derivative

$$f'_\varepsilon(x) = -\varepsilon e^{-\varepsilon x} + \varepsilon e^{-x} > 0$$

when $x > 0$ and $\varepsilon > 1$, so $f_\varepsilon(x) > 0$ when $x > 0$. ◀

The following result is a common technique for proving sub-exponential.

▶ **Proposition 33.** *For any random variable X with mean μ and any number λ, $\mathbb{E}\left[e^{X-\mu}\right] < \mathbb{E}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right]$ where X' is a random variable which is independent and identical to X.*

Proof.

$$\mathbb{E}_X\left[e^{\lambda(X-\mu)}\right] = \mathbb{E}_X\left[e^{\lambda(X-\mathbb{E}_{X'}[X'])}\right] \le \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]$$

The second inequality is due to Jensens inequality. Let ε be a random variable taking value on ± 1 with probability half on both values. Since X and X' are identical, $\varepsilon(X-X')$ and $X - X'$ are identical. So we have

$$\mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_\varepsilon\left[e^{\varepsilon\lambda(X-X')}\right]\right]$$

On the other hand, for any number Y,

$$\mathbb{E}_\varepsilon\left[e^{\varepsilon Y}\right] = \frac{1}{2}(e^Y + e^{-Y}) = \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{Y^k}{k!} + \frac{(-Y)^k}{k!}\right)$$

$$= \sum_{k=1}^{\infty}\left(\frac{Y^{2k}}{(2k)!}\right) < \sum_{k=1}^{\infty}\left(\frac{Y^{2k}}{2^k k!}\right) = e^{Y^2/2}$$

So $\mathbb{E}_{X,X'}\left[\mathbb{E}_\varepsilon\left[e^{\varepsilon\lambda(X-X')}\right]\right] < \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right]$ ◀

A.2 Useful Bounds

In this section, we review some concentration or anti-concentration bounds which we will use later.

▶ **Proposition 34.** *Let $X = x_1 + x_2 + \cdots + x_{m'}$ be the sum of m' i.i.d Bernoulli numbers with probability $\frac{c\cdot A}{n}$. Let $\hat{A} = \frac{Xn}{cm'}$. Then the probability that $\left|\hat{A} - A\right| \le \delta_0$ is $O(n^{-2.5})$ if $m' > \frac{10A}{c\delta_0^2} n \log n$.*

Proof. By Chernoff bound, for any $0 < \epsilon < 1$,

$$\Pr[|X - \frac{m'cA}{n}| > \frac{\epsilon m'cA}{n}|] < e^{-\frac{\epsilon^2 m'cA}{4n}}$$

Let $\epsilon = \frac{c\delta_0}{A}$, the RHS will be $e^{-\frac{\delta_0^2 m}{4cn}} < e^{-2.5\log n} = O(n^{-2.5})$, ◀

▶ **Definition 35** (Sub-exponential Variables)**.** *A random variable X with mean μ is sub-exponential with parameters (σ, b) if for any λ with $|\lambda| < 1/b$,*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\sigma^2\lambda^2/2}$$

▶ **Proposition 36** (Bernstein bound [2]). *Let $X_1, X_2, \ldots, X_n$ be independent random variables, where X_i is sub-exponential random variable with mean μ_i and sub-exponential parameter (σ_i, b_i).*

$$\Pr\left(\left|\sum_{i=1}^{n}(X_i-\mu_i)\right|\geq t\right)\leq\begin{cases}2e^{-\frac{t^2}{2\sigma_\star^2}} & \textit{for } 0\leq t\leq\frac{\sigma_\star}{b}\\ 2e^{-\frac{t}{2b_\star}} & \textit{for } t>\frac{\sigma_\star}{b}\end{cases}$$

where $\sigma_\star^2=\sum_{i=1}^n\sigma_i^2$ and $b_\star=\max_{i=1}^n b_i$

▶ **Definition 37** (Lévy Concentration Function [7]). *Given a random variable X and a number t, the Lévy Concentration function $Q_X(t)$ is defined as*

$$Q_X(t)=\sup_{a\in\mathbb{R}}\Pr\left(|X-a|<t\right)$$

▶ **Proposition 38** (Kolmogorov-Rogozin Inequality [10]). *Let $X_1, X_2, \ldots, X_n$ be independent random variables and let $X=X_1+X_2+\cdots+X_n$. Then for any $t>0$ and any $0<t_i<t$, we have*

$$Q_X(t)\leq 100\frac{t}{\sqrt{\sum_{i=1}^n t_i^2(1-Q_{X_i}(t_i))}}$$

B Omitted Details from Section 2.2

▷ Lemma (Restatement of Lemma 10). For any vertex v_i which located at position x_i on the segment, if we uniformly sample a vertex v on the segment, then the edge (v_i, v) will be present with probability $\frac{c}{n}(2-e^{-x_i}-e^{x_i-n})$.

Proof. The probability is the expectation of $e^{-|x_i-x|}$ where x is the location of v which is uniformly sampled on the segment. So the probability is

$$\begin{aligned}&\int_0^n\frac{c}{n}e^{-|x_i-x|}dx\\ =&\frac{c}{n}\int_0^{x_i}e^{x-x_i}dx+\frac{c}{n}\int_{x_i}^n e^{x_i-x}dx\\ =&\frac{c(2-e^{-x_i}-e^{x_i-n})}{n}\end{aligned}$$

◀

▷ Lemma (Restatement of Lemma 11). For any two vertices v_i and v_j which are located at x_i and x_j on the segment with $x_i<x_j$, if we uniformly sample a vertex v on the segment, then v is a common neighbor of v_i and v_j with probability $\frac{c^2}{n}((x_j-x_i+1)e^{x_i-x_j}-\frac{1}{2}(e^{x_i+x_j-2n}+e^{-x_i-x_j}))$.

Proof. Let $p(x)$ be the probability that v is a common neighbor of v_i and v_j where x is the location of v, then

$$p(x)=\begin{cases}c^2\cdot e^{2x-x_i-x_j}, & \text{if } x\leq x_i\\ c^2\cdot e^{x_i-x_j}, & \text{if } x_i<x<x_j\\ c^2\cdot e^{x_i+x_j-2x}, & \text{if } x\geq x_j\end{cases}$$

So the overall probability is

$$\begin{aligned}
&\int_0^n \frac{1}{n} p(x)dx \\
=&\frac{c^2}{n}\int_0^{x_i} e^{2x-x_i-x_j}dx + \frac{c^2(x_j-x_i)}{n}e^{x_i-x_j} + \frac{c^2}{n}\int_{x_j}^{n} e^{x_i+x_j-2x}dx \\
=&\frac{c^2(x_j-x_i+1)}{n}e^{x_i-x_j} - \frac{c^2(e^{-x_i-x_j}+e^{x_i+x_j-2n})}{2n}
\end{aligned}$$

◀

▷ Lemma (Restatement of Lemma 12). If $m > \frac{2500n\log n}{c^2\delta^2}$, with probability $1-o(1)$, for any two vertices v_i and v_j, (a) if they have no common neighbor, then $|x_i - x_j| > 2.5$, and (b) if $|x_i - x_j| > n/2$, then they have no common neighbor.

Proof. If $|x_i - x_j| \leq 2.5$, then one of $e^{-x_i-x_j}$ and $e^{x_i+x_j-2n}$ is $O(e^{-n})$, without loss of generality, suppose $e^{x_i+x_j-2n}$ is $O(e^{-n})$. Since $-x_i - x_j < -|x_i - x_j|$, $e^{-x_i-x_j} < e^{-|x_i-x_j|}$. By Lemma 11, the probability that a random sampled vertex be a common neighbor of v_i and v_j is at least $\frac{c^2(|x_i-x_j|+0.5)}{n}e^{-|x_i-x_j|} > \frac{c^2}{2n}e^{-2.5} > \frac{c^2}{30n}$. Since $m > \frac{2500n\log n}{c^2\delta^2}$, the probability that v_i and v_j have no common neighbor is $o(n^{-80})$.

If $|x_i - x_j| > n/2$, the probability that a random vertex be a common neighbor of them is at most $e^{-n/2}$. So with probaiblity $1 - o(n^{-100})$, they have no common neighbor. ◀

C Distance Approximation for Inverse Linear Decaying Function

In this section, we deal with the case that $f(x) = \frac{c}{x+1}$ and thus the probability of an edge existing between two vertex v_i and v_j with location x_i and x_j on the segment be $\frac{c}{|x_i-x_j|+1}$. We first analyze the degree of each vertex and the number of common neighbors between each two vertices.

▶ **Lemma 39.** *Suppose a vertex v_i is located at x_i, if we uniformly sample a vertex v on the segment then an edge (v_i, v) will be presented with probability* $\frac{c\log(x_i+1)+c\log(n-x_i+1)}{n}$

Proof. The probability is

$$\begin{aligned}
\frac{c}{n}\int_0^n (|x-x_i|+1)^{-1}dx &= \frac{c}{n}(\int_1^{x_i+1} x^{-1}dx + \int_1^{n-x_i+1} x^{-1}dx) \\
&= \frac{c(\log(x_i+1)+\log(n-x_i+1))}{n}
\end{aligned}$$

◀

▶ **Lemma 40.** *Suppose two vertices v_i and v_j are located at x_i and x_j on the segment with $x_i < x_j$ and $d = x_j - x_i$, if we uniformly sample a vertex v on the segment, then v is a common neighbor of v_i and v_j with probability*

$$\frac{c^2}{n}\left(\log(d+1)\left(\frac{2}{d}+\frac{2}{d+2}\right)+\frac{1}{d}(\log(x_i+1)-\log(x_j+1)+\log(n-x_j+1)-\log(n-x_i+1))\right)$$

Proof. The probability is

$$
\begin{aligned}
&\frac{c^2}{n}\int_0^n (|x-x_i|+1)^{-1}(|x-x_j|+1)^{-1}dx\\
=&\frac{c^2}{n}\left(\int_1^{x_i+1}\frac{1}{x(x+d)}dx+\int_1^{d+1}\frac{1}{x(d+2-x)}dx+\int_1^{n-x_j+1}\frac{1}{x(x+d)}dx\right)\\
=&\frac{c^2}{n}\left(\int_1^{x_i+1}\frac{1}{d}\left(\frac{1}{x}-\frac{1}{(x+d)}\right)dx+\int_1^{n-x_j+1}\frac{1}{d}\left(\frac{1}{x}-\frac{1}{x+d}\right)dx+\int_1^{d+1}\frac{1}{d+2}\left(\frac{1}{x}+\frac{1}{(d+2-x)}\right)dx\right)\\
=&\frac{c^2}{n}\left(\frac{1}{d}(\log(x_i+1)-\log(x_j+1)+\log(n-x_j+1)-\log(n-x_i+1)+2\log(d+1))+\frac{1}{d+2}(2\log(d+1))\right)
\end{aligned}
$$

◀

Then, we prove that we can check if a vertex v_i is close to one of the endpoints. If so, we can further approximate its location with a multiplicative error. In the rest of this section, let $\varepsilon=\frac{\delta}{20}$.

▶ **Lemma 41.** *If $m>\frac{40n\log^2}{c\varepsilon^2}$ and $0<\varepsilon<\frac{1}{10}$, then with probability $1-o(1)$, for any vertex v_i, we can output a number $\hat{x}_i$ such that:*

- *If $\bar{x}_i>\frac{9}{\varepsilon}-1$, then $\hat{x}_i>\frac{2}{\varepsilon}+1$.*
- *If $\bar{x}_i\le\frac{9}{\varepsilon}-1$, then $|\hat{x}_i-\bar{x}_i|<(1+\varepsilon)(\bar{x}_i+1)$.*

where $\bar{x}_i=\min\{x_i,n-x_i\}$.

Proof. Since $m>\frac{100n\log^2 n}{c\varepsilon^2}$. By Proposition 34 and Lemma 39, we can approximate $\log(x_i+1)+\log(n-x_i+1)=\log(\bar{x}_i+1)+\log(n-\bar{x}_i+1)$ within additive error $\frac{\varepsilon}{3}$ with probability $1-o(1)$. Let a be this value, we prove that $\hat{x}_i=e^{a-\log n}-1$ satisfies the requirement.

$a-\log n=\log(\frac{(\bar{x}_i+1)(n-\bar{x}_i+1)}{n})\pm\frac{\varepsilon}{3}=\log(\bar{x}_i+1)+\log(1-\frac{\bar{x}_i-1}{n})\pm\frac{\varepsilon}{3}$. By Proposition 29, $\log(1-\frac{\bar{x}-1}{n})=o(1)$ if $\bar{x}_i<\frac{9}{\varepsilon}-1$ and at most 1 otherwise.

If $\bar{x}_i>\frac{9}{\varepsilon}-1$, $a-\log n>\log(\frac{9}{\varepsilon})-1-\frac{\varepsilon}{3}>\log(\frac{3}{\varepsilon})-\frac{\varepsilon}{3}$. So $\hat{x}_i>(1-\frac{\varepsilon}{2})\cdot\frac{3}{\varepsilon}-1=\frac{3}{\varepsilon}-2.5>\frac{2}{\varepsilon}+1$ since $\varepsilon<\frac{1}{10}$.

If $\bar{x}_i\le\frac{9}{\varepsilon}-1$, $a-\log n=\log(\bar{x}_i+1)\pm\frac{\varepsilon}{2}$ So $\hat{x}_i+1=(1\pm(e^{\varepsilon/2}))(\bar{x}_i+1)=(1\pm\varepsilon)(\bar{x}_i+1)$. ◀

▶ **Lemma 42.** *Suppose $0<\delta<0.1$ and $m>\frac{16000n\log^2 n}{c\delta^2}$, with probability $1-o(1)$, for any two vertex v_i and v_j with distance d, we can approximate d by $\hat{d}$ which satisfies:*

- *$\hat{d}<d+\delta$ if $d<0.3$.*
- *$d-\delta<\hat{d}<d+\delta$ if $0.3\le d\le 2$.*
- *$\hat{d}>d-\delta$ if $d>2$.*

Proof. For any number a, b, denote $g(a,b)=\frac{\log a-\log b}{a-b}$ and $h(a)=\log(a+1)(\frac{2}{a}+\frac{2}{a+2})$. We first prove that we can either approximate $h(d)$ with additive error at most 2ε or directly output a $\hat{d}$ which satisifies the condition. By Lemma 40 and Proposition 34, we can approximate $h(d)-g(x_i+1,x_j+1)-g(n-x_i+1,n-x_j+1)$ with additive error $\frac{\varepsilon}{c\sqrt{\log n}}=o(1)$, Denote a as this value.

Let $\hat{x}_i$ and $\hat{x}_j$ be the value given by Lemma 41. If $\hat{x}_i$ and $\hat{x}_j$ are both at least $\frac{1}{\varepsilon}$, then v_i and v_j are both at least $\frac{1}{\varepsilon}-1$ far away from both endpoints. By the argument in the proof of Proposition 28, $g(x_i+1,x_j+1)$ and $g(n-x_i+1,n-x_j+1)$ are both at most ε. So $|a-h(d)|<2\varepsilon$. If one of $\hat{x}_i$ and $\hat{x}_j$ larger than $\frac{2}{\varepsilon}+1$ and the other less than $\frac{1}{\varepsilon}$, then $|x_j-x_i|>\frac{2}{\varepsilon}-(1+\varepsilon)\frac{1+\varepsilon}{\varepsilon}>2$. So we can directly output $\hat{d}=n$. The only case remaining is when both of $\hat{x}_i$ and $\hat{x}_j$ at most $\frac{2}{\varepsilon}+1$.

In this case, x_i and x_j are both at most $\frac{3}{\varepsilon}$ far away from one of the endpoint. If they are close to different endpoint, then $d > n/2$, which menas $\mathbb{E}[a] = O(\frac{1}{n})$ and $a = o(1)$. Otherwise $\mathbb{E}[a] = \Omega(1) - o(1)$ and thus $a = \Omega(1)$. So we can check if v_i and v_j are close to the same endpoint. If not, $x_j - x_i > n/2$ and so we can directly output $\hat{d} = n$. Then we focus on the case that they are close to the same endpoint. Without loss of generality, suppose both of x_i and x_j are at most $\frac{3}{\varepsilon}$.

If $\hat{x}_i$ and $\hat{x}_j$ are both at most 8, then both of x_i and x_j are at most $9(1+\varepsilon)-1 < 9$, which means $|\hat{x}_i - x_i|$ and $|\hat{x}_j - x_j|$ are both at most $10\varepsilon = \frac{\delta}{2}$. Then we can output $\hat{d} = |\hat{x}_i - \hat{x}_j|$. If one of $\hat{x}_i$ and $\hat{x}_j$ is at least 8 and the other is at most 5, then $|x_i - x_j| > 3(1-2\varepsilon) > 2$. So we can output $\hat{d} = n$. The only case remaining is when both of $\hat{x}_i$ and $\hat{x}_j$ are at least 5. In this case, x_i and x_j are both larger than 4. By Proposition 28, $|g(x_i, x_j) - g(\hat{x}_i, \hat{x}_j)| < \varepsilon$. So $a - g(\hat{x}_i, \hat{x}_j)$ is an approximation of $h(d)$ with additive error at most $\varepsilon + o(1) < 2\varepsilon$.

By this point, we either already output a $\hat{d}$ which satisfies the condition or have an approximation of $h(d)$ with additive error 2ε. To complete the proof we observe that the function $h(d)$ is monotone decreasing when $d > 0$ and that the derivative of $h(d)$ is strictly less than -0.1 when $0.5 \le d \le 2$. ◀

Note that if $0 < \delta < 0.1$, $3\delta < 0.5 < \frac{n}{2}$ and $2 > 0.5 + 8\delta$. Theorem 3 immediately follows from Lemma 5 and Lemma 42.

D Omitted Details from Section 3.1

▷ Lemma (Restatement of Lemma 17). If $|x_i - x_j| \ge n^{0.1}$, then $|L_{i,j}| < n^{-100}$.

Proof. By definition of Y, $d'_{i,j} = \frac{2\delta d_{i,j}}{n}$. So

$$L_{i,j} \le \log(1-e^{-d_{i,j}}) - \log(1-e^{d'_{i,j}-d_{i,j}}) = \log(1 - \frac{e^{d'_{i,j}-d_{i,j}}(1-e^{-d'_{i,j}})}{1-e^{-d_{i,j}}}) < \log(1 - \frac{e^{d'_{i,j}-d_{i,j}}}{1-e^{-d_{i,j}}})$$

Since $d_{i,j} \ge n^{0.1}$, $e^{-d_{i,j}}$ and $e^{d'_{i,j}-d_{i,j}}$ are both $o(n^{-100})$. So $L_{i,j} = \log(1 - o(n^{-100})) = o(n^{-100})$ by Proposition 29. ◀

▷ Lemma (Restatement of Lemma 18). For any pair of vertices v_i, v_j, $e^{-d_{i,j}}(d'^2_{i,j}/2 + (1 - e^{-d_{i,j}})a/2) < \mathbb{E}[L_{i,j}] < e^{-d_{i,j}}(d'^2_{i,j} + \frac{2d'^2_{i,j}}{d_{i,j}})$ where $a = \frac{e^{-d_{i,j}}(e^{d'_{i,j}}-1)}{1-e^{-d_{i,j}}}$.

Proof. By definition of $L_{i,j}$, with probability $e^{-d_{i,j}}$, $L_{i,j} = -d'_{i,j}$ and with probability $1 - e^{-d_{i,j}}$, $L_{i,j} = \log(1-e^{-d_{i,j}}) - \log(1-e^{-d_{i,j}+d'_{i,j}}) = -\log(\frac{1-e^{-d_{i,j}+d'_{i,j}}}{1-e^{-d_{i,j}}})$. So

$$\begin{aligned}\mathbb{E}[L_{i,j}] &= -d'_{i,j}e^{-d_{i,j}} - (1-e^{-d_{i,j}})\log(\frac{1-e^{-d_{i,j}+d'_{i,j}}}{1-e^{-d_{i,j}}}) \\ &= -d'_{i,j}e^{-d_{i,j}} - (1-e^{-d_{i,j}})\log(1 - \frac{e^{-d_{i,j}}(e^{d'_{i,j}}-1)}{1-e^{-d_{i,j}}})\end{aligned}$$

by Proposition 30, $a = \frac{e^{-d_{i,j}}(e^{d'_{i,j}}-1)}{1-e^{-d_{i,j}}} < \frac{d'_{i,j}}{d_{i,j}} < 0.5$. Together with Proposition 29,

$$\mathbb{E}[L_{i,j}] < -d'_{i,j}e^{-d_{i,j}} + e^{-d_{i,j}}(e^{d'_{i,j}}-1)(1+a) < e^{-d_{i,j}}(e^{d'_{i,j}} - d'_{i,j} - 1 + \frac{d'_{i,j}(e^{d'_{i,j}}-1)}{d_{i,j}})$$

Since $d'_{i,j} < 1/2$, $e^{d'_{i,j}} < 1 + d'_{i,j} + d'^2_{i,j}$ and $e^{d'_{i,j}} < 1 + 2d'_{i,j}$. So $\mathbb{E}[L_{i,j}] < e^{-d_{i,j}}(d'^2_{i,j} + 2d'^2_{i,j}/d_{i,j})$.

Again by Proposition 29,

$$\begin{aligned}\mathbb{E}[L_{i,j}] &> -d'_{i,j}e^{-d_{i,j}} + e^{-d_{i,j}}(e^{d'_{i,j}} - 1)(1 + a/2)\\ &> e^{-d_{i,j}}(e^{d'_{i,j}} - d'_{i,j} - 1 + (e^{d'_{i,j}} - 1)a/2)\\ &> e^{-d_{i,j}}(d'^2_{i,j}/2 + (e^{d'_{i,j}} - 1)a/2)\end{aligned}$$

◀

▷ Lemma (Restatement of Lemma 19). If $m < \frac{(10^{-5})n^{3/2}}{\delta}$ and X is obtained by sampling each point uniformly, then $\mathbb{E}\left[\sum_{i,j} L_{i,j}\right] < 10^{-8}$ with probability $1 - o(1)$.

Proof. Let the $S_1, S_2, \ldots, S_n$ be the set of vertices where S_k contains all the vertices inside the interval $[i, i+1]$ in X. Let i, j be two vertices inside S_k and S_ℓ where $k \leq \ell$, then $\mathbb{E}[L_{i,j}] \leq 6(\ell - k + 1)^2 e^{-(\ell-k-1)} \cdot \frac{\delta^2}{n^2}$ by Lemma 18 and the fact that the distance between i and j is at least $\ell - k - 1$ and at most $\ell - k + 1$, $|y_i - y_j| = (1 - \frac{2\delta}{n})|x_i - x_j|$. So

$$\begin{aligned}\mathbb{E}\left[\sum_{i,j} L_{i,j}\right] &= \sum_{k,\ell}\sum_{i\in S_k, j\in S_\ell}\mathbb{E}[L_{i,j}]\\ &\leq \frac{\delta^2}{n^2}\sum_{k,\ell}|S_k|\cdot|S_\ell|6(\ell-k+1)^2e^{-(\ell-k-1)}\\ &= \frac{\delta^2}{n^2}\sum_{k=0}^{n-1}\sum_{\ell=1}^{n-k}|S_\ell|\cdot|S_{\ell+k}|6(k+1)^2e^{-(k-1)}\end{aligned}$$

By Rearrangement inequality [12], for any k, $\sum_{\ell=1}^{n-k}|S_\ell|\cdot|S_{\ell+k}| \leq \sum_{\ell=1}^{n}|S_\ell|^2$. So

$$\begin{aligned}\mathbb{E}\left[\sum_{i,j} L_{i,j}\right] &\leq \frac{\delta^2}{n^2}\Big(\sum_{k=1}^{n}|S_k|^2\Big)\cdot\Big(\sum_{k=0}^{n-1}6(k+1)^2e^{-(k-1)}\Big)\\ &\leq \frac{\delta^2}{n^2}\Big(6e + \sum_{k=0}^{\infty}(6k^2+24k+24)e^{-k}\Big)\Big(\sum_{k=1}^{n}|S_k|^2\Big)\\ &\leq \frac{\delta^2}{n^2}\Big(6e + \frac{6e(1+e)}{(e-1)^3} + \frac{24e}{(e-1)^2} + \frac{24e}{e-1}\Big)\cdot\Big(\sum_{k=1}^{n}|s_k|^2\Big)\\ &\leq \frac{100\delta^2}{n^2}\sum_{k=1}^{n}|S_k|^2\end{aligned}$$

By the choice of m, each $|S_k| < 2m/n < \frac{10^{-5}n^{1/2}}{\delta}$ with probability $1 - o(1)$ by Chernoff bound, so $\sum_{k=1}^{n}|S_k|^2 \leq \frac{10^{-10}n^2}{\delta^2}$, which means $\mathbb{E}\left[\sum_{i,j} L_{i,j}\right] < 10^{-8}$. ◀

▷ Lemma (Restatement of Lemma 20). $\Pr\left(\left|\bar{L} - \mathbb{E}[\bar{L}]\right| > 10^{-3}\sqrt{\mathbb{E}[\bar{L}]}\right) \geq 0.5$.

Proof. For any i and j, let $t_{i,j} = (d'_{i,j} - \log(\frac{1-e^{d'_{i,j}-d_{i,j}}}{1-e^{-d_{i,j}}}))/2$. We prove that $t^2_{i,j}(1 - Q_{L_{i,j}}(t_{i,j})) \geq \frac{1}{20}\mathbb{E}[L_{i,j}]$ where $Q_{L_{i,j}}$ is the Lèvy concentration function of $L_{i,j}$ (see Definition 37).

Sincee $L_{i,j}$ is either $-d'_{i,j}$ or $-\log(\frac{1-e^{d'_{i,j}-d_{i,j}}}{1-e^{-d_{i,j}}})$, $Q_{L_{i,j}}(t_{i,j})$ is the maximum between $e^{-d_{i,j}}$ and $1 - e^{-d_{i,j}}$. Also, $2t_{i,j}$ is larger than both $d'_{i,j}$ and $\log(\frac{1-e^{d'_{i,j}-d_{i,j}}}{1-e^{-d_{i,j}}})$.

If $e^{-d_{i,j}} \leq 1/2$, $d_{i,j} \geq \log 2$, $t_{i,j}^2(1-Q_{L_{i,j}}(t_{i,j})) \geq \frac{1}{4}(d_{i,j}'^2 e^{-d_{i,j}})$, which is larger than $\frac{1}{20}(1+2/d_{i,j})d_{i,j}'^2 e^{-d_{i,j}} > \frac{1}{16}\,\mathbb{E}\left[L_{i,j}\right]$ sincee $d_{i,j} > \log 2$ and Lemma 18.

Otherwise $e^{-d_{i,j}} > 1/2$, $t_{i,j}^2(1-Q_{L_{i,j}}(t_{i,j})) \geq \frac{1}{4}((1-e^{-d_{i,j}})\log^2(\frac{1-e^{d_{i,j}'-d_{i,j}}}{1-e^{-d_{i,j}}})$. Let $a = \frac{e^{-d_{i,j}}(e^{d_{i,j}'}-1)}{1-e^{-d_{i,j}}}$, By Lemma 18, $\mathbb{E}\left[L_{i,j}\right] = -d_{i,j}' e^{-d_{i,j}} - (1-e^{-d_{i,j}})\log(1-a) > 0$, so $-(1-e^{-d_{i,j}})\log(1-a) > d_{i,j}' e^{-d_{i,j}}$, which means

$$(1-e^{-d_{i,j}})\log^2(\frac{1-e^{d_{i,j}'-d_{i,j}}}{1-e^{-d_{i,j}}}) = (1-e^{-d_{i,j}})\log^2(1-a) > -d_{i,j}' e^{-d_{i,j}}\log(1-a) > a d_{i,j}' e^{-d_{i,j}}$$

Since $e^{-d_{i,j}} > 1/2 > 1-e^{-d_{i,j}}$, $a = \frac{e^{-d_{i,j}}(e^{d_{i,j}'}-1)}{1-e^{-d_{i,j}}} > e^{d_{i,j}'}-1 > d_{i,j}'$, also $a = \frac{e^{-d_{i,j}}(e^{d_{i,j}'}-1)}{1-e^{-d_{i,j}}} > \frac{e^{d_{i,j}'}-1}{2(1-e^{-d_{i,j}})} > \frac{d_{i,j}'}{2d_{i,j}}$. So $5a > d_{i,j}' + 2d_{i,j}'/d_{i,j}$, which means

$$t_{i,j}^2(1-Q_{L_{i,j}}(t_{i,j})) \geq \frac{1}{4}d_{i,j}' e^{-d_{i,j}} a > \frac{1}{20}d_{i,j}'^2(1+2/d_{i,j})e^{-d_{i,j}} > \frac{1}{20}\,\mathbb{E}\left[L_{i,j}\right]$$

by Lemma 18.

By Proposition 38,

$$Q_{\bar{L}}(10^{-3}\sqrt{\mathbb{E}\left[\bar{L}\right]}) \leq \frac{10^{-1}\sqrt{\mathbb{E}\left[\bar{L}\right]}}{\sqrt{\sum_{i\sim j} t_{i,j}^2 Q_{L_{i,j}}(t_{i,j})}} \leq \frac{0.1\sqrt{\mathbb{E}\left[\bar{L}\right]}}{\sqrt{0.05\sum_{i\sim j}\mathbb{E}\left[L_{i,j}\right]}} \leq 0.5$$

◀

▷ **Lemma (Restatement of Lemma 21).** For any pair of i and j, if $d_{i,j} < n^{0.1}$, then $L_{i,j}$ is a sub-exponential variable with parameters $(\sigma_{i,j}, b)$ where $\sigma_{i,j}^2 = 48\,\mathbb{E}\left[L_{i,j}\right]$ and $b = n^{-0.8}$.

Proof. By Proposition 33, it is sufficient to prove for any $|\lambda| < n^{0.8}$,

$$\mathbb{E}_{L_{i,j},L_{i,j}'}\left[e^{\frac{\lambda^2(L_{i,j}-L_{i,j}')^2}{2}}\right] < e^{\frac{\lambda^2\sigma_{i,j}^2}{2}}$$

Where $L_{i,j}'$ is a random variable independent and identical to $L_{i,j}$.

Again, by convenience, denote $a = \frac{e^{-d_{i,j}}(e^{d_{i,j}'}-1)}{1-e^{-d_{i,j}}}$. We first bound $|L_{i,j}|$, $L_{i,j}$ is either $-d_{i,j}'$ or $-\log(1-a)$ where $d_{i,j}' = \frac{2d_{i,j}\delta}{n} = o(n^{-0.8})$. Also by Proposition 29 and Proposition 30, $-\log(1-a) < a + a^2 < \frac{d_{i,j}'}{d_{i,j}} + \frac{d_{i,j}'^2}{d_{i,j}^2} < \frac{2\delta}{n} = o(n^{-0.8})$. So $L_{i,j} - L_{i,j}' = o(n^0.8)$. So $\lambda^2(L_{i,j}-L_{i,j}')^2 = o(1)$ for any $|\lambda| < n^{0.8}$. So

$$\mathbb{E}_{L_{i,j},L_{i,j}'}\left[e^{\frac{\lambda^2(L_{i,j}-L_{i,j}')^2}{2}}\right] < 1 + \mathbb{E}_{L_{i,j},L_{i,j}'}\left[\lambda^2(L_{i,j}-L_{i,j}')^2\right]$$

On the other hand,

$$e^{\frac{\lambda^2\sigma_{i,j}^2}{2}} > 1 + \frac{\lambda^2\sigma_{i,j}^2}{2}$$

To prove the lemma, it's sufficient to prove $\sigma_{i,j}^2 > 2\mathbb{E}_{L_{i,j},L_{i,j}'}\left[(L_{i,j}-L_{i,j}')^2\right]$. By definition of $L_{i,j}$,

$$\begin{aligned}\mathbb{E}_{L_{i,j},L_{i,j}'}\left[(L_{i,j},L_{i,j}')^2\right] &= 2e^{-d_{i,j}}(1-e^{-d_{i,j}})(d_{i,j}'-\log(1-a))^2\\ &< 4e^{-d_{i,j}}d_{i,j}'^2 + 4(1-e^{-d_{i,j}})\log^2(1-a)\end{aligned}$$

By Proposition 29 and Proposition 30, $a < 1/2$ and $\log^2(1-a) < (a+a^2)^2 < 3a^2$. So

$$4e^{-d_{i,j}} d'^2_{i,j} + 4(1-e^{-d_{i,j}})\log^2(1-a) < 4e^{-d_{i,j}}(d'^2_{i,j} + 3(e^{d'_{i,j}}-1)a) < 24\,\mathbb{E}\left[L_{i,j}\right]$$

by Lemma 18. The proof finish with $\sigma^2_{i,j} = 48\,\mathbb{E}\left[L_{i,j}\right]$. ◀

▷ Lemma (Restatment of Lemma 22). If $1000n < m < \frac{(10^{-5})n^{3/2}}{\delta}$, then $\mathbb{E}\left[\bar{L}\right] = \omega(n^{-1.6})$.

Proof. By Chernoff bound, if $m > 1000n$, with probability $1-o(1)$, there is a vertex in any segments with length $\frac{\log n}{20}$. Without loss of generality, suppose $x_1 < x_2 < \cdots < x_m$, then $x_{i+1} - x_i < \frac{\log n}{10}$ for any i and $x_m - x_1 > n - \log n > 0.9n$. So by Lemma 18,

$$\mathbb{E}\left[\bar{L}\right] > \sum_{i=1}^{m-1} \mathbb{E}\left[L_{i,i+1}\right] > e^{-\frac{\log n}{20}} \sum_{i=1}^{m-1} (\frac{2\delta(x_{i+1}-x_i)}{n})^2 = \omega(n^{-2.1}) \sum_{i=1}^{m-1}(x_{i+1}-x_i)^2$$

By Cauchy-Schwarz inequality,

$$\sum_{i=1}^{m-1}(x_{i+1}-x_i)^2 > \frac{1}{m}(\sum_{i=1}^{m-1}(x_{i+1}-x_i))^2 > \frac{1}{m}(0.9n)^2 = \Omega(n^{0.5})$$

So $\mathbb{E}\left[\bar{L}\right] = \omega(n^{-1.6})$. ◀

▷ Lemma (Restatment of Lemma 23). If $1000n < m < \frac{(10^{-5})n^{3/2}}{\delta}$, then for any integer $k > 0$, $\Pr\left(\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| > 10k\sqrt{\mathbb{E}\left[\bar{L}\right]}\right) < 4e^{-k}$.

Proof. By Proposition 36 and Lemma 21,

$$\Pr\left(\left|\bar{L} - \mathbb{E}\left[\bar{L}\right]\right| > 10k\sqrt{\mathbb{E}\left[\bar{L}\right]}\right) < 2e^{-\frac{100k^2\,\mathbb{E}[\bar{L}]}{2\sigma^2_\star}} + 2e^{-\frac{10k\sqrt{\mathbb{E}[\bar{L}]}}{2b}}$$

where $\sigma^2_\star = \sum_{i\ j}\sigma^2_{i,j} = 48\,\mathbb{E}\left[\bar{L}\right]$ and $b = n^{-0.8} < \sqrt{\mathbb{E}\left[\bar{L}\right]}$ by Lemma 22. So

$$2e^{-\frac{100k^2\,\mathbb{E}[\bar{L}]}{2\sigma^2_\star}} + 2e^{-\frac{10k\sqrt{\mathbb{E}[\bar{L}]}}{2b}} < e^{-\frac{100k^2}{96}} + 2e^{-\frac{10k}{2}} < 4e^{-k}$$

◀

E Omitted Details from Section 3.2

▷ Lemma (Restatement of Lemma 24). If $m = \tilde{O}(n^2)$, with probability $1-o(1)$, for any pair of (i,j), $|x_i - x_j| > \frac{1}{n^4}$.

Proof. For any pair of (i,j), the probability that $|x_i - x_j| \le \frac{1}{n^4}$ is at most $\frac{\frac{2}{n^4}}{n} = O(\frac{1}{n^5})$. Since there are at most $m^2 = o(n^5)$ pair of (i,j), so with probability $1-o(1)$ there is no such (i,j). ◀

▷ Lemma (Restatement of Lemma 25). For those pair of (i,j) contribute to $\bar{L}$, $\bar{L}_{i,j}$ is a sub-exponential random variable with parameter $(\sigma_{i,j}, b)$ where $\sigma^2_{i,j} = 10\log n \cdot \mathbb{E}\left[\bar{L}_{i,j}\right]$ and $b = 10\log n$.

Proof. By Proposition 33, it is sufficient to prove that for any $\lambda < \frac{1}{b}$,

$$\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[e^{\frac{\lambda^2(\bar{L}_{i,j}-\bar{L}'_{i,j})^2}{2}}\right] < e^{\frac{\lambda^2\sigma^2_{i,j}}{2}}$$

where $\bar{L}'_{i,j}$ is a random varibale independent and identical to $\bar{L}_{i,j}$. We prove the lemma respectively in the case of $|y_i - y_j| \leq |x_i - x_j|$ and $|y - i - y_j| < |x_i - x_j|$.

Case 1: $|y_i - y_j| \leq |x_i - x_j|$. Denote $a = \frac{e^{-d_{i,j}}(e^{d'_{i,j}}-1)}{1-e^{-d_{i,j}}}$. $\bar{L}_{i,j} = -d'_{i,j}$ with probability $e^{-d_{i,j}}$ and $a + \frac{1}{2}a^2$ with probability $(1 - e^{-d_{i,j}})$. So

$$\begin{aligned}
\mathbb{E}\left[\bar{L}_{i,j}\right] &= -d'_{i,j}e^{-d_{i,j}} + (1 - e^{-d_{i,j}})(a + \frac{1}{2}a^2) \\
&= e^{-d_{i,j}}(e^{d'_{i,j}} - d_{i,j} - 1) + \frac{1}{2}(1 - e^{-d_{i,j}})a^2 \\
&\geq \frac{1}{2}(e^{-d_{i,j}}d'^2_{i,j} + (1 - e^{-d_{i,j}})a^2)
\end{aligned}$$

So $e^{\frac{\lambda^2\sigma^2}{2}} > 1 + 5\lambda^2 e^{-d_{i,j}}(d'^2_{i,j} + a^2)\log n$.

On the other hand, $(\bar{L}_{i,j} - \bar{L}'_{i,j})^2 = (d'_{i,j} + a)^2 \leq 2d'^2_{i,j} + 2a^2$ with probability $2e^{-d_{i,j}}(1 - e^{-d_{i,j}}$ and 0 otherwise. By the condition that $\bar{L}_{i,j}$ contributes to $\bar{L}$, $d'_{i,j} \leq d_{i,j} \leq 5\log n$; by Proposition 30, $a \leq 1$. So $\lambda^2(2d'^2_{i,j} + 2a^2) < \frac{50\log^2 n+2}{100\log^2 n} < 1$ for any $\lambda < \frac{1}{b}$. Which means

$$\begin{aligned}
\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[e^{\frac{\lambda^2(\bar{L}_{i,j}-\bar{L}'_{i,j})^2}{2}}\right] &\leq 1 + \mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[\lambda^2(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] \\
&\leq 1 + 2e^{-d_{i,j}}(1 - e^{-d_{i,j}})(2d'^2_{i,j} + 2a^2)\lambda^2
\end{aligned}$$

which means

$$\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[e^{\frac{\lambda^2(\bar{L}_{i,j}-\bar{L}'_{i,j})^2}{2}}\right] < e^{\frac{\lambda^2\sigma^2_{i,j}}{2}}$$

Case 2: $|y_i - y_j| > |x_i - x_j|$. Denote $a = \log(\frac{1-e^{-(d_{i,j}+d'_{i,j})}}{1-e^{-d_{i,j}}})$. $\bar{L}_{i,j} = \min\{d'_{i,j}, 2\}$ with probability $e^{-d_{i,j}}$ and $-a$ with probability $(1 - e^{-d_{i,j}})$. Since $d_{i,j} \geq \frac{1}{n^4}$, $\log(1 - e^{-d_{i,j}}) \geq \log(1 - e^{-n^{-4}}) \geq \log(\frac{1}{2n^4}) \geq -5\log n$, which means $a < 5\log n$. So $\lambda^2(\bar{L}_{i,j} - \bar{L}'_{i,j})^2 \leq (2a^2 + 2)\lambda^2 < 1$ for any $\lambda < \frac{1}{10\log n} = \frac{1}{b}$. So

$$\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[e^{\frac{\lambda^2(\bar{L}_{i,j}-\bar{L}'_{i,j})^2}{2}}\right] \leq 1 + \mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[\lambda^2(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right]$$

On the other hand, $e^{\frac{\lambda^2\sigma^2}{2}} > 1 + 5\lambda^2\,\mathbb{E}\left[\bar{L}_{i,j}\right]\log n$, so we just need to prove

$$\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] \leq 5\,\mathbb{E}\left[\bar{L}_{i,j}\right]\log n$$

Case 2.1: If $d'_{i,j} \geq 2$,

$$\begin{aligned}
\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] &= 2e^{-d_{i,j}}(1 - e^{-d_{i,j}})(8 + 2a^2) < 16e^{-d_{i,j}} + 4(1 - e^{-d_{i,j}})a\log n \\
&< 16e^{-d_{i,j}} + 4e^{-d_{i,j}}(1 - e^{-d'_{i,j}})\log n < 5e^{-d_{i,j}}\log n
\end{aligned}$$

On the other hand, $\mathbb{E}\left[\bar{L}_{i,j}\right] = 2e^{-d_{i,j}} - (1 - e^{-d_{i,j}})a > e^{-d_{i,j}}(2 - (1 - e^{-d'_{i,j}})) > e^{-d_{i,j}}$, so $5\,\mathbb{E}\left[\bar{L}_{i,j}\right]\log n > 5e^{-d_{i,j}}\log n$.

Case 2.2: If $d'_{i,j} \leq d_{i,j}$ and $d'_{i,j} < 2$, $\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] < 2e^{-d_{i,j}}(1-e^{-d_{i,j}})(2d'^2_{i,j} + 2a^2)$. Let $z = \frac{e^{-d_{i,j}}(1-e^{-d'_{i,j}})}{1-e^{-d_{i,j}}}$, by Proposition 31, $z < \frac{d'_{i,j}}{d_{i,j}} \leq 1$, so $a = \log(1+z) < z - \frac{z^2}{2} + \frac{z^3}{3} < z - \frac{z^2}{6}$, which means $\mathbb{E}\left[\bar{L}_{i,j}\right] > e^{-d_{i,j}}d'_{i,j} - (1-e^{-d_{i,j}})(z - \frac{z^2}{6}) = e^{-d_{i,j}}(d'_{i,j} + e^{-d'_{i,j}} - 1) + \frac{1}{6}z^2(1-e^{-d_{i,j}})$ where $e^{-d'_{i,j}} + d'_{i,j} - 1 > \frac{d'^2_{i,j}}{2} - \frac{d'^3_{i,j}}{6} > \frac{d'^2_{i,j}}{6}$ since $d'_{i,j} < 2$. So $\mathbb{E}\left[\bar{L}_{i,j}\right] > \frac{1}{6}e^{-d_{i,j}}d'^2_{i,j} + \frac{1}{6}z^2(1-e^{-d_{i,j}})$. On the other hand, $2e^{-d_{i,j}}(1-e^{-d_{i,j}})(2d'^2_{i,j} + 2a^2) < 4e^{-d_{i,j}}d'^2_{i,j} + 4(1-e^{-d_{i,j}})z^2$. So $\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] < \mathbb{E}\left[\bar{L}_{i,j}\right] \cdot \log n$.

Case 2.3: If $d_{i,j} < d'_{i,j} < 2$, let $\varepsilon = e^{-d_{i,j}}\frac{d'_{i,j}}{d_{i,j}}$ and $z = \frac{e^{-d_{i,j}}(1-e^{-d'_{i,j}})}{1-e^{-d_{i,j}}}$. Since $a = \log(1+z) < z$, $\mathbb{E}\left[\bar{L}_{i,j}\right] > e^{-d_{i,j}}(d'_{i,j} - 1 + e^{-d'_{i,j}}) > \frac{1}{6}e^{-d_{i,j}}d'^2_{i,j}$ since $d'_{i,j} < 2$. On the other hand, since $d'_{i,j} > d_{i,j}$, $\frac{1-e^{-d'_{i,j}}}{1-e^{-d_{i,j}}} < \frac{d'_{i,j}}{d_{i,j}}$ by Proposition 32, so $a < \log(1 + e^{-d_{i,j}}\frac{d'_{i,j}}{d_{i,j}}) = \log(1+\varepsilon)$, and $\mathbb{E}\left[\bar{L}_{i,j}\right] > d_{i,j}\varepsilon - (1-e^{-d_{i,j}})\log(1+\varepsilon) > d_{i,j}(\varepsilon - \log(1+\varepsilon)) > \frac{d_{i,j}}{2}\log(1+\varepsilon)^2$ (the last inequality is due to $e^a - a - 1 > \frac{a^2}{2}$ for any $a > 0$). So

$$\mathbb{E}\left[\bar{L}_{i,j}\right] > \frac{1}{2}(\frac{1}{6}e^{-d_{i,j}}d'^2_{i,j} + \frac{d_{i,j}}{2}\log(1+\varepsilon)^2) > \frac{1}{24}(e^{-d_{i,j}}(1-e^{-d_{i,j}})(d'^2_{i,j} + 3a^2))$$

So

$$\mathbb{E}_{\bar{L}_{i,j},\bar{L}'_{i,j}}\left[(\bar{L}_{i,j} - \bar{L}'_{i,j})^2\right] \leq 4e^{-d_{i,j}}(1-e^{-d_{i,j}})(d'^2_{i,j} + a^2) \leq 5\,\mathbb{E}\left[\bar{L}_{i,j}\right]\log n$$

◀

The following lemma shows that $d'_{i,j}$ satisfies the triangle inequality.

▶ **Lemma 43.** *For any i, j and k, $d'_{i,j} \leq d'_{i,k} + d'_{k,j}$.*

Proof. Since X and Y has the same vertex order, $d'_{i,k} + d'_{k,j} = |x_i - x_k - y_i + x_k| + |x_j - x_k - y_j + y_k| \geq |x_i - x_j - y_i + y_j| = d'_{i,j}$. ◀

▷ Lemma (Restatement of Lemma 27). If $\frac{100n^{1.5}\log n}{\delta} < m = \Omega(n^2)$. If X is sampled uniformly, then with probability $1 - o(1)$, for any Y such that there is a pair of i and j satisfies that $d'_{i,j} > \frac{\delta}{2}$, $\mathbb{E}\left[\bar{L}\right] > 5\log^2 n$.

Proof. By Lemma 26, $\mathbb{E}\left[\bar{L}_{i,j}\right] \geq 0$. It is sufficient to prove that sum of some $\mathbb{E}\left[\bar{L}_{i,j}\right]$ contributed to $\bar{L}$ is larger than $5\log n$. We first prove that if there is a pair of i' and j' satisfies $d_{i',j'} \leq 1$ and $d'_{i',j'} > \frac{\delta}{8}$, then $\mathbb{E}\left[\bar{L}\right] > 5\log^2 n$. By Chernoff bound, with probability $1 - o(1)$ there are at least $\frac{90\sqrt{n}\log n}{\delta}$ vertices in each segment of length 1. So there are at least $\frac{90\sqrt{n}\log n}{\delta}$ vertices which is at most 1 away from both $v_{i'}$ and $v_{j'}$. Suppose v_k is such a vertex, then either $d'_{i',k}$ or $d'_{k,j'}$ is at least $\frac{\delta}{16}$ by Lemma 43, which means either $\mathbb{E}\left[\bar{L}_{i',k}\right]$ or $\mathbb{E}\left[\bar{L}_{k,j'}\right]$ is at least $\frac{\delta^2}{256e}$ by lemma 26. So $\bar{L} > \frac{90\sqrt{n}\log n}{\delta} \cdot \frac{\delta^2}{256e} > 5\log^2 n$.

For any integer K, let S_K be the set of vertex in segment $[K-1, K]$. Let $v_i \in S_I$ and $v_j \in S_J$. Without loss of generality, suppose $I \leq J$. Then for any vertex v_k in S_I (resp. S_j), if $d'_{i,k}$ (resp. $d'_{k,j}$) is at least $\frac{\delta}{8}$, which means $\mathbb{E}\left[L\right] > 5\log^2 n$. Otherwise, we have $I < J$ and for any $v_k \in S_I$ and $v_\ell \in S_J$, $d'_{k,\ell} > \frac{\delta}{4}$.

For any $I \leq K \leq J$, let v_{k_K} be an arbitrary vertex in S_K. We prove that $\sum_{I \leq K < J}\mathbb{E}\left[\bar{L}_{k_K,k_{K+1}}\right] \geq \frac{\delta^2}{1000n}$. For any K, since v_{k_K} and $v_{k_{K+1}}$ are in S_K and S_{K+1} respectively, $d_{k_K,k_{K+1}} \leq 2$, which means $e^{-d_{k_K,k_{K+1}}} > e^{-2} > \frac{1}{10}$. If there exists a K such that $d'_{k_K,k_{K+1}} > 2$, then $\bar{L}_{k_K,k_{K+1}} > \frac{1}{10} > \frac{\delta^2}{1000n}$ by Lemma 27. Otherwise $\sum_{I \leq K < J}\mathbb{E}\left[\bar{L}_{k_K,k_{K+1}}\right] \geq \frac{1}{60}\sum_{I \leq K < J} d'^2_{k_K,k_{k+1}}$ by

Lemma 27. Since $d'_{k_I,k_J} > \frac{\delta}{4}$, $\sum_{I \le K < j} d'_{k_K,k_{k+1}} \ge \frac{\delta}{4}$ by Lemma 43. By Cauchy-Schwarz inequality,

$$\sum_{I \le K < j} d'^2_{k_K,k_{k+1}} \ge \frac{1}{J-I}\Big(\sum_{I \le K < j} d'_{k_K,k_{k+1}}\Big)^2 \ge \frac{\delta^2}{16(J-I)} \ge \frac{\delta^2}{16n}$$

which means $\sum_{I \le K < J} \mathbb{E}\left[\bar{L}_{k_K,k_{K+1}}\right] \ge \frac{\delta^2}{1000n}$.

Let $N = \frac{90\sqrt{n}}{\delta}$ and for any $I \le K \le J$, let $v_{\ell_1^K}, v_{\ell_2^K}, \ldots, v_{\ell_N^K}$ be arbitrary N vertices in S_k. Then

$$\begin{aligned}
\mathbb{E}\left[\bar{L}\right] &\ge \sum_{I \le K < J} \sum_{i'=1}^{N} \sum_{j'=1}^{N} \mathbb{E}\left[\bar{L}_{\ell_{i'}^K,\ell_{j'}^{K+1}}\right] = \sum_{I \le K < J} \sum_{i'=1}^{N} \sum_{j'=0}^{N-1} \mathbb{E}\left[\bar{L}_{\ell_{i'}^K,\ell_{(i'+j') \bmod N+1}^{K+1}}\right] \\
&= \sum_{i'=1}^{N} \sum_{j'=0}^{N-1} \sum_{I \le K < J} \mathbb{E}\left[\bar{L}_{\ell^K_{(i'+Kj') \bmod N+1},\ell^{K+1}_{(i'+(K+1)j') \bmod N+1}}\right] \\
&\ge \sum_{i'=1}^{N} \sum_{j'=0}^{N-1} \frac{\delta^2}{1000n} = \frac{N^2\delta^2}{1000n} > 5\log^2 n
\end{aligned}$$

◀