Seeded Graph Matching via Large Neighborhood Statistics

Seeded Graph Matching via Large Neighborhood Statistics

Elchanan Mossel
MIT
elmos@mit.edu

Jiaming Xu
Duke University
jiaming.xu868@duke.edu

July 27, 2018

Abstract

We study a well known noisy model of the graph isomorphism problem. In this model, the goal is to perfectly recover the vertex correspondence between two edge-correlated graphs, with an initial seed set of correctly matched vertex pairs revealed as side information. Specifically, the model first generates a parent graph G_0 from Erdős-Rényi random graph $\mathcal{G}(n,p)$ and then obtains two children graphs G_1 and G_2 by subsampling the edge set of G_0 twice independently with probability $s = \Theta(1)$. The vertex correspondence between G_1 and G_2 is obscured by randomly permuting the vertex labels of G_1 according to a latent permutation π^*. Finally, for each i, $\pi^*(i)$ is revealed independently with probability α as seeds.

In the sparse graph regime where $np \le n^\epsilon$ for any $\epsilon < 1/6$, we give a polynomial-time algorithm which perfectly recovers π^*, provided that $nps^2 - \log n \to +\infty$ and $\alpha \ge n^{-1+3\epsilon}$. This further leads to a sub-exponential-time, $\exp\left(n^{O(\epsilon)}\right)$, matching algorithm even without seeds. On the contrary, if $nps^2 - \log n = O(1)$, then perfect recovery is information-theoretically impossible as long as α is bounded away from 1.

In the dense graph regime, where $np = bn^a$, for fixed constants $a, b \in (0,1]$, we give a polynomial-time algorithm which succeeds when $b = O(s)$ and $\alpha = \Omega\left((np)^{-\lfloor 1/a \rfloor} \log n\right)$. In particular, when $a = 1/k$ for an integer $k \ge 1$, $\alpha = \Omega(\log n/n)$ suffices, yielding a quasi-polynomial-time $n^{O(\log n)}$ algorithm matching the best known algorithm by Barak et al. for the problem of graph matching without seeds when $k \ge 153$ and extending their result to new values of p for $k = 2, \ldots, 152$.

Unlike previous work on graph matching, which used small neighborhoods or small subgraphs with a logarithmic number of vertices in order to match vertices, our algorithms match vertices if their large neighborhoods have a significant overlap in the number of seeds.

1 Introduction

In this paper, we study a well-known model of noisy graph isomorphism. Our main interest is in polynomial time algorithms for seeded problems where the matching between a small subset of the nodes is revealed. For seeded problems, our result provides a dramatic improvement over previously known results. Our results also shed light on the unseeded problem. In particular, we give (the first) sub-exponential time algorithms for sparse models and an $n^{O(\log n)}$ algorithm for dense models for some parameters, including some that are not covered by recent results of Barak et al. [BCL+18].

We recall that two graphs are isomorphic if there exists an edge-preserving bijection between their vertex sets. The Graph Isomorphism problem is not known to be solvable in polynomial time, except in special cases such as graphs of bounded degree [Luk80] and bounded eigenvalue multiplicity [BGM82]. However, a recent breakthrough of Babai [Bab16] gave a quasi-polynomial time algorithm.

In a number of applications including network security [NS09, NS08], systems biology [SXB08], computer vision [CFSV04, SS05], and natural language processing [HNM05], we are given two graphs as input which we believe have an underlying isomorphism between them. However, they are not exactly isomorphic because they have each been perturbed in some way, adding or deleting edges randomly. This suggests a noisy version of Graph Isomorphism also known as *graph matching* [LR13], where we seek a bijection that minimizes the number of edge disagreements.

Given two graphs with adjacency matrices G_1 and G_2, if our goal is to minimize the ℓ_2 distance between G_1 and some permuted version of G_2, then graph matching can be viewed as a special case of the *quadratic assignment problem* (QAP) [BCPP98]: namely,

$$\min_{\Pi} \|G_1 - \Pi G_2 \Pi^\top\|_F^2, \tag{1}$$

where Π ranges over all $n \times n$ permutation matrices, and $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ denotes the Frobenius norm. QAP is NP-hard in the worst case. There are exact search methods for QAP based on branch-and-bound and cutting planes, as well as various approximation algorithms based on linearization schemes, and convex/semidefinite programming relaxations (see [FQRM+16] and the references therein). However, approximating QAP within a factor $2^{\log^{1-\epsilon}(n)}$ for $\epsilon > 0$ is NP-hard [MMS10].

These hardness results only apply in the worst case, where the two graphs are designed by an adversary. However, in many aforementioned applications, we are not interested in worst-case instances, but rather in instances for which there is enough information in the data to recover the underlying isomorphism, i.e., when the amount of data or signal-to-noise ratio is above the information-theoretic limit. The key question is whether there exists an efficient algorithm that is successful all the way down to this limit. In this vein, we consider the following random graph model denoted by $\mathcal{G}(n, p; s)$ [PG11].

Definition 1 (The Correlated Erdős-Rényi model $\mathcal{G}(n, p; s)$)**.** *Suppose we generate a parent graph G_0 from the Erdős-Rényi random graph model $\mathcal{G}(n, p)$. For a fixed realization of G_0, we generate two subgraphs G_1 and G_2 by subsampling the edges of G_0 twice. More specifically,*

- *We let G_1^* be a random subgraph of G_0 obtained by including every edge of G_0 with probability s independently.*
- *We repeat the above subsampling procedure, but independently to obtain another random subgraph of G_0, denoted by G_2.*

To further model the scenario that we do not know the vertex correspondence between G_1 and G_2 a prior, we sample a random permutation π^* over $[n]$ and let G_1 denote the graph obtained by relabeling every vertex i in G_1^* as $\pi^*(i)$.

The goal is to exactly recover π^* from the observation of G_1 and G_2 with high probability, i.e., to design an estimator $\widehat{\pi}$ based on G_1 and G_2 such that

$$\mathbb{P}\left\{\widehat{\pi}(G_1, G_2) = \pi^*\right\} \to 1, \quad \text{as } n \to \infty.$$

As a motivating example, we can model G_0 as some true underlying friendship network of n persons, G_1 is an anonymized Facebook network of the same set of persons, and G_2 is a Twitter network with known person identities. If we can recover the vertex correspondence between G_1 and G_2, then we can de-anonymize the Facebook network G_1 (this example ignores many important facts such as additional graph structures in real life networks).

Note that s is equal to the probability of $e \in E(G_2)$ conditional on $e \in E(G_1)$, and hence can be viewed as a measure of the edge correlations. Throughout this paper, without further specifications, we shall assume $s = \Theta(1)$.

In the fully sampling case $s = 1$, graph matching under $\mathcal{G}(n, p; 1)$ reduces to the Graph Automorphism problem for Erdős-Rényi graphs. In this case, a celebrated result [Wri71] shows that if $\log n + \omega(1) \le np \le n - \log n - \omega(1)$, then with probability $1 - o(1)$, the size of the automorphism group of G_0 is 1 and hence the underlying permutation π^* can be exactly recovered; otherwise, with probability $1 - o(1)$, the size of the automorphism group of G_0 is strictly bigger than 1 and hence exact recovery of the underlying permutation is information-theoretically impossible. Recent work [CK16, CK17][1] has extended this result to the partially sampling case $s = \Theta(1)$ and $p \le 1/2$, showing that the Maximum Likelihood Estimator, or equivalently the optimum of QAP (1), coincides with the ground truth π^* with high probability, provided that $nps^2 \ge \log n + \omega(1)$; on the contrary, any estimator is correct with probability $o(1)$, if $nps^2 \le \log n - \omega(1)$.

From a computational perspective, in the fully sampling case $s = 1$, there exist linear-time algorithms which attain the recovery threshold, in the sense that they exactly recover the underlying permutation with high probability whenever $np = \log n + \omega(1)$ [Bol82, CP08]. However, in the partially sampling case, it is still open whether any efficient algorithm can succeed close to the threshold. A recent breakthrough result [BCL+18] obtains a quasi-polynomial-time ($n^{O(\log n)}$) algorithm which succeeds when $np \ge n^{o(1)}$ and $s \ge (\log n)^{-o(1)}$. However, this is still far away from the information-theoretic limit $nps^2 \ge \log n + \omega(1)$.

Another line of work [PG11, YG13, KL14, LFP13, FAP18, SGE17] in this area considers a relaxed version of the graph matching problem, where an initial seed set of correctly matched vertex pairs is revealed as side information. This is motivated by the fact that in many real applications, some side information on the vertex identities are available and have been successfully utilized to match many real-world networks [NS09, NS08]. Formally, in this paper, we assume the seed set is randomly generated as follows.

Definition 2 (Seeded graph matching under $\mathcal{G}(n, p; s, \alpha)$)**.** *In addition to G_1, G_2 that are generated under $\mathcal{G}(n, p; s)$ with a latent permutation π^*, we have access to π_0 such that $\pi_0(i) = \pi^*(i)$ with probability α and $\pi_0(i) =?$ with probability $1 - \alpha$ independently across different i. The goal is to recover π^* based on G_1, G_2, and π_0.*

The vertex i such that $\pi_0(i) = \pi^*(i)$ is called seeded vertices and the set of seed vertices is denoted by I_0. Note that according to our model, the number of seeds $|I_0|$ is distributed as

[1] In fact, a more general correlated Erdős-Rényi random graph model is considered in [CK16, CK17], where $\mathbb{P}\{G_1(i,j) = a, G_2(i,j) = b\} = p_{a,b}$ for $a, b \in \{0, 1\}$.

Binom(n, α). For a given size K, we could also consider a deterministic size model where I_0 is chosen uniformly at random from all possible subsets of $[n]$ with size K. The main results of this paper readily extend to this deterministic size model with $K = \lfloor n\alpha \rfloor$.

The results of the seeded graph matching turn out to be useful for designing graph matching without seeds. On the one hand, when a seed set of size K is not given, we could obtain it in $n^{O(K)}$ steps by randomly choosing a set of K vertices and then enumerating all the possible mapping. This is known as the beacon set approach to graph isomorphism [Lip78]. On the other hand, we could first apply a seedless graph matching algorithm and then apply a seeded graph matching algorithm to boost its accuracy. This two-step algorithms have been successful both theoretically [BES80] [Bol01, Section 3.5] and empirically [LFP13].

In the sparse graph regime $np = \Theta(\log n)$, it is shown in [YG13] that if $\alpha = \Omega(1/\log^2 n)$, or equivalently, the size of the seed set is $\Omega(n/\log^2 n)$, then a percolation-based graph matching algorithm correctly matches $n - o(n)$ vertices in polynomial-time with high probability. In the dense graph regime $np = n^\delta$ for some constant $\delta \in (0, 1)$, a seed set of size $\Theta(n^{1-\delta})$ suffices as shown in [YG13]. Another work [KL14] shows that if $nps^2\alpha \geq 24 \log n$, then one can match all vertices correctly in polynomial-time with high probability based on counting the number of "common" seeded vertices. Note that this exact recovery result requires the seed set size to be linear in n in the sparse graph regime $np = \Theta(\log n)$.

In summary, despite a significant amount of previous work on seedless and seeded graph matching, the following two fundamental questions remain elusive:

Question 1. *In terms of graph sparsity, can we achieve the information-theoretic limit* $nps^2 - \log n \to +\infty$ *in sub-exponential, or polynomial time?*

Question 2. *In terms of seed set, what is the minimum number of seeds required for exact recovery in sub-exponential, or polynomial time?*

Our main results shed light on this two questions by improving the state-of-the-art of seeded graph matching. First, we show that it is possible to achieve the information theoretic limit $nps^2 \geq \log n + \omega(1)$ of graph sparsity in polynomial-time. Then, we show the number of seeds needed for exact recovery in polynomial-time can be as low as n^ϵ in the sparse graph regime ($np \leq n^\epsilon$) and $\Omega(\log n)$ in the dense graph regime.

1.1 Main Results

We first consider the sparse graph regime.

Theorem 1. *Suppose* $np \leq n^{1/2-\epsilon}$ *for a fixed constant* $\epsilon > 0$ *and* $s = \Theta(1)$. *Assume*

$$nps^2 - \log n \to +\infty \tag{2}$$

$$\alpha \geq n^{-1/2+3\epsilon}. \tag{3}$$

Then there exists a polynomial-time algorithm, namely Algorithm 1, which outputs $\widehat{\pi} = \pi^*$ *with probability at least* $1 - o(1)$ *under the seeded* $\mathcal{G}(n, p; s, \alpha)$ *model.*

Notice that (4) is the information-theoretic limit for graph matching under the seedless $\mathcal{G}(n, p; s)$ model. In fact, Theorem 2 shows that (4) is necessary for seeded graph matching as long as α is bounded away from 1. Its proof is standard and can be found in Appendix A.

Theorem 2. *If*

$$nps^2 - \log n = O(1),$$

then any algorithm outputs $\widehat{\pi} \neq \pi^*$ with at least a probability of $\Omega\left((1-\alpha)^2\right)$ under the seeded $\mathcal{G}(n, p; s, \alpha)$ model.

Also, the condition (5) requires that the size of the seed set is $n^{1/2+3\epsilon}$ compared to the best previously known results that required the seed set to be almost linear in n.

It is natural to ask if $n^{1/2}$ seeded nodes are required for polynomial time algorithm. While from the proof of Theorem 1, it might look that $n^{1/2}$ is optimal due to the birthday paradox effect, it turns out we can do better!

The following result relaxes the size of seed set needed to $n^{3\epsilon}$.

Theorem 3. *Suppose $np \leq n^{\epsilon}$ for a fixed constant $\epsilon < 1/6$ and $s = \Theta(1)$. Assume*

$$nps^2 - \log n \to +\infty \tag{4}$$

$$\alpha \geq n^{-1+3\epsilon}. \tag{5}$$

Then there exists a polynomial-time algorithm, namely Algorithm 3, which outputs $\widehat{\pi} = \pi^*$ with probability at least $1 - o(1)$ under the seeded $\mathcal{G}(n, p; s, \alpha)$ model.

We next consider the dense graph regime, where we assume the average degree np is parameterized as:

$$np = bn^a \tag{6}$$

for some fixed constants $a, b \in (0, 1]$. Let

$$d = \left\lfloor \frac{1}{a} \right\rfloor + 1, \tag{7}$$

Theorem 4. *Consider the dense graph regime (6). Assume*

$$b \leq \frac{s}{16(2-s)^2}, \tag{8}$$

and

$$\alpha \geq \frac{300 \log n}{(nps^2)^{d-1}}, \tag{9}$$

where d is given in (7). Then there exists an polynomial-time algorithm, namely Algorithm 2, which outputs $\widehat{\pi} = \pi^*$ with probability $1 - 4n^{-1}$ under the seeded $\mathcal{G}(n, p; s, \alpha)$ model.

Our results for seeded graph matching also imply the results for graph matching without seeds.

Theorem 5. *Suppose a Seeded Graph Matching algorithm outputs $\widehat{\pi} = \pi^*$ with high probability under the seeded graph matching model $\mathcal{G}(n, p; s, \alpha)$. Assume $nps^2 - \log n \to +\infty$ and $\alpha n \to +\infty$. Then there exists an algorithm, namely Algorithm 4, which calls the Seeded Graph Matching algorithm $n^{O(\alpha n)}$ times and outputs $\widehat{\pi} = \pi^*$ under the seedless model $\mathcal{G}(n, p; s)$ with high probability.*

Remark 1. *Consider the dense regime (6) with $a = 1/k$ for an integer $k \geq 1$. Then $d = k+1$ and $(np)^{d-1} = b^k n$. Hence, as shown by Theorem 4, $\alpha n \geq 300 \log n (bs^2)^{-k}$, or equivalently $\Omega(\log n)$ number of seeds, suffice for exact recovery in polynomial-time. Since we can enumerate over all possible matchings for $\log n$ seeds in quasi-polynomial $n^{O(\log n)}$ time, this implies a quasi-polynomial time matching algorithm even without seeds, as shown by Theorem 5. The previous work [BCL+18] gives a quasi-polynomial time matching algorithm in the range*

$$np \in \left[n^{o(1)}, n^{1/153}\right] \cup \left[n^{2/3}, n^{1-\epsilon}\right].$$

Our results complement their results by filling in gaps in the above range with points $np \in \{bn^{1/k} : 1 \leq k \leq 152\}$.

1.2 Key Algorithmic Ideas and Analysis Techniques

Most previous work [PG11, YG13, KL14, LFP13, FAP18, SGE17] on seeded graph matching exploits the seeded information by looking at the number of seeded vertices that are *direct* neighbors of a given vertex. Since the average degree of a vertex is np, $np\alpha \gg 1$ is needed so that there are sufficiently many seeded vertices that are direct neighbors of a given vertex.

Our idea is to explore much bigger ("global") neighborhoods of a given vertex up to radius ℓ for a suitably chosen ℓ, and match two vertices by comparing the set of seeded vertices in their ℓ-th local neighborhoods. This idea was used before in the noiseless and seedless case, in [Bol82, CP08] but to the best of our knowledge was not used in the noisy and seeded case. Since we are looking at global neighborhoods, we can only perform very simple tests. Indeed, the test we perform to check if two vertices are matched is just to count how many seeded vertices do the two neighborhoods have in common. Thus, our algorithms are very simple.

The main challenge in the analysis is to control the size of neighborhoods of the coupled graphs G_0, G_1 and G_2. In this regard, we draw on a number of tools from the literature on studying subgraph counts [JLR11] and the diameter in random graphs [Bol01]. See Appendix D for details.

2 Our Algorithms

Before presenting our algorithms, we first explain why (4) is needed for graph matching under $\mathcal{G}(n, p, s)$. Denote the intersection graph and the union graph by $G_1^* \wedge G_2$ and $G_1^* \vee G_2$. Then

$$G_1^* \wedge G_2 \sim \mathcal{G}(n, ps^2) \quad \text{and} \quad G_1^* \vee G_2 \sim \mathcal{G}(n, ps(2-s)).$$

Notice that $G_1^* \wedge G_2$ contains the statistical signature for matching vertices, as a subgraph in $G_1^* \wedge G_2$ will appear in both G_1 and G_2. If $nps^2 - \log n = O(1)$, then classical random graph theory implies that with high probability, $G_1^* \wedge G_2$ contains isolated vertices. The underlying true vertex correspondence of these isolated vertices cannot be correctly matched; hence the impossibility of exact recovery. See Appendix A for a precise argument.

In contrast, if $nps^2 - \log n \to +\infty$, then $G_1^* \wedge G_2$ is connected with high probability. Moreover, for a high-degree vertex i in $G_1^* \wedge G_2$, its local neighbhorhood grows like a branching process. In particular, the number of vertices at distance ℓ from i is approximately $(nps^2)^\ell$. Furthermore, for a pair of two vertices i, j chosen at random in $G_1^* \vee G_2$, the intersection of the local neighborhoods of i and j is typically of size $O\left((nps)^{2\ell} n^{-1}\right)$. Therefore, if $(nps^2)^\ell \gg (nps)^{2\ell} n^{-1}$ and $\alpha(nps^2)^\ell \gg 1$, a large number of vertices can be distinguished with high probability based on the set of seeded vertices in their ℓ-th local neighborhoods. This is the key idea underlying our algorithms.

We shall use the following notations of local neighborhoods. For a given graph G, we denote by $\Gamma_k^G(u)$ the set of vertices at distance k from v in G:

$$\Gamma_k^G(u) = \{v \in V(G) : d(u,v) = k\} \tag{10}$$

and write $N_k^G(u)$ for the set of vertices within distance k from u:

$$N_k^G(u) = \cup_{i=0}^k \Gamma_i(u). \tag{11}$$

When the context is clear, we abbreviate $\Gamma_k^G(u)$ and $N_k^G(u)$ as $\Gamma_k(u)$ and $N_k(u)$ for simplicity.

2.1 A Simple Algorithm in Sparse Graph Regime

We first present a simple seeded graph matching algorithm which succeeds up to the information-theoretic limit in terms of graph sparsity when the initial seed set is of size $n^{1/2+3\epsilon}$.

The idea of the algorithm is based on matching ℓ-th local neighborhoods of two vertices by finding independent paths (vertex-disjoint except for the starting vertex) to seeded vertices. The ℓ is chosen such that $(np)^\ell \approx n^{1/2-\epsilon}$. In this setting, we expect that if i in G_1 and j in G_2 are true matches, then their local neighborhoods intersect a lot; if i and j are wrong matches, then their local neighborhoods barely intersect. Hence, if $\alpha(nps^2)^\ell \gg 1$, then we can find a sufficiently large number of, say m, independent (vertex-disjoint except for i) paths of length ℓ from i to m seeded vertices in $\Gamma_\ell^{G_1^* \wedge G_2}(i)$. Such m paths of length ℓ form a starlike tree T in $G_1^* \wedge G_2$ with root vertex i and a set of m seeded leaves, denoted by L (See Fig. 1 for an example of $m = 3$ and $\ell = 2$). Note that T will appear in G_2 with root vertex i and the set of seeded leaves L; it will also appear in G_1 with root vertex $\pi^*(i)$ and the corresponding set of seeded leaves $\pi^*(L)$. However, since the ℓ-th local neighborhoods of two distinct vertices barely intersect, T will *not* appear in $G_1^* \vee G_2$ with a root vertex other than i. Therefore, we can correctly match the vertex $\pi^*(i)$ in G_1 with the high-degree vertex i in G_2 by finding such a starlike tree T, or equivalently m independent ℓ-paths to a set of m common seeded vertices.

Algorithm 1 Graph matching based on counting independent ℓ-paths to seeded vertices

1: **Input:** G_1, G_2, π_0, $m, \ell \in \mathbb{Z}$
2: **Output:** $\widehat{\pi}$.
3: **Match high-degree vertices:** For each pair of unseeded vertices $i_1 \in V(G_1)$ and $i_2 \in V(G_2)$, if there are m independent ℓ-paths in G_2 from i_2 to a set of m seeded vertices $L \subset \Gamma_\ell^{G_2}(i_2)$, and there are m independent ℓ-paths in G_1 from i_1 to the corresponding set of m seeded vertices $\pi_0(L) \subset \Gamma_\ell^{G_1}(i_1)$, then set $\widehat{\pi}(i_2) = i_1$. Declare failure if there is any conflict.
4: **Match low-degree vertices:** For every $i_2 \in I_0$, set $\widehat{\pi}(i_2) = \pi_0(i_2)$. For all the pairs of unmatched vertices (i_1, i_2), if i_1 is adjacent to a matched vertex j_1 in G_1 and i_2 is adjacent to vertex $\widehat{\pi}(j_1)$ in G_2, set $\widehat{\pi}(i_2) = i_1$. Declare failure if there is any conflict.
5: Output $\widehat{\pi}$ to be a random permutation when failure is declared or there is any vertex unmatched.

There are two tuning parameters ℓ and m in Algorithm 1. Later in our analysis, we will optimally choose

$$\ell = \left\lfloor \left(\frac{1}{2} - \epsilon\right) \frac{\log n}{\log(nps^2)} \right\rfloor \geq 1 \tag{12}$$

and

$$m = \left\lceil \frac{2}{\epsilon} \right\rceil. \tag{13}$$

Note that when $nps^2 - \log n \to +\infty$, there may exist vertices with small degrees. In fact, classical random graph results say that the minimum degree of $\mathcal{G}(n,p)$ is k with high probability for a fixed integer k, provided that

$$(k-1)\log\log n + \omega(1) \le nps^2 - \log n \le k\log\log n - \omega(1),$$

see, e.g., [FK15, Section 4.2]. Hence, due to the existence of low-degree vertices, we may not be able to match all vertices correctly at one time based on the number of independent paths to seeded vertices. Our idea is to first match high-degree vertices and then match the remaining low-degree vertices with the aid of high-degree vertices matched in the first step. In particular, we let

$$\tau = \frac{nps^2}{\log(nps^2)}. \tag{14}$$

We say a vertex i high-degree, if its degree $d_i \ge \tau$ in $G_1^* \wedge G_2$; otherwise, we say it is a low-degree vertex. As we will see in Section 3, conditioning on that $G_1^* \wedge G_2$ and $G_1^* \vee G_2$ satisfy some typical graph properties, all low-degree vertices can be easily matched correctly given a correct matching of high-degree vertices.

In passing, we remark on the time complexity of Algorithm 1. Note that for ease of presentation, in Algorithm 1, we do not specify how to efficiently find out whether there exist m independent ℓ-paths in G_2 from i_2 to seed set $L \subset \Gamma_\ell^{G_2}(i_2)$, and m independent ℓ-paths in G_1 from i_1 to the corresponding seed set $\pi_0(L) \subset \Gamma_\ell^{G_1}(i_1)$. It turns out for a given pair of vertices i_1, i_2, this task can be reduced to a maximum flow problem in a directed graph, which can be solved via Ford–Fulkerson algorithm [FF56] in $O(n\alpha)$ time steps (see Appendix E for details). Since there are at most n^2 pairs of vertices i_1, i_2 to consider, Step 3 of Algorithm 1 taks at most $O(n^3\alpha)$. The Step 4 of matching low-degree vertices in Algorithm 1 takes at most $O(n^3 p)$ time steps. Hence, in total Algorithm 1 takes at most $O\left(n^3(\alpha+p)\right)$ time steps.

2.2 A Simple Algorithm in Dense Graph Regime

In this subsection, we consider the dense graph regime given in (6), where $np = bn^a$ and $d = \lfloor 1/a \rfloor + 1$. In this setting, since $p^d n^{d-1} - 2\log n \to +\infty$ and $p^{d-1}n^{d-2} - 2\log n \le -\infty$, it follows from [Bol01, Corollary 10.12] that $\mathcal{G}(n,p)$ has diameter d with high probability. Thus, when $s = \Theta(1)$, both $G_1^* \wedge G_2$ and $G_1^* \vee G_2$ have diameter d with high probability. Therefore, we present an algorithm based on matching the $d-1$-th local neighborhood of two vertices. More specifically, our algorithm matches $i_1 \in V(G_1)$ and vertex $i_2 \in V(G_2)$ based on the number of seeded vertices *within* distance $d-1$ from i_1 in G_1 and *within* distance $d-1$ from i_2 in G_2.

Algorithm 2 Graph matching based on $(d-1)$-hop witness in dense regime

1: **Input:** G_1, G_2, π_0, $d \in \mathbb{Z}$.
2: **Output:** $\widehat{\pi}$.
3: **Match all vertices:** For each pair of unseeded vertices $i_1 \in V(G_1)$ and $i_2 \in V(G_2)$, compute

$$w_{i_1,i_2} = \left|\left\{ j \in I_0 : \pi_0(j) \in N_{d-1}^{G_1}(i_1),\ j \in N_{d-1}^{G_2}(i_2) \right\}\right|. \tag{15}$$

Set $\widehat{\pi}(i_2) \in \arg\max_{i_1} w_{i_1,i_2}$. Set $\widehat{\pi}(i_2) = \pi_0(i_2)$ for each seeded vertex $i_2 \in I_0$. Declare failure if there is any conflict.

Algorithm 2 runs in polynomial-time. The precise running time depends on the data structures for storing and processing graphs. To be specific, let us assume it takes one time step to fetch the

set of direct neighbors of a given vertex. Then fetching the set $N_\ell^G(i)$ of all vertices within distance ℓ from a given vertex i takes a total of $O(|N_\ell^G(i)|) = O(n)$ time steps. Thus computing w_{i_1,i_2} in (15) for a given pair of vertices i_1, i_2 takes at most $O(n)$ time steps. Hence, in total Algorithm 2 takes $O(n^3)$ time steps. One could possibly obtain a better running time via a more careful analysis or a better data structure.

The difference in the analysis compared to the first algorithm is that the $(d-1)$-th local neighborhoods are not tree-like anymore. Instead, we have to analyze the exposure process of the two neighborhoods, for which we use a previous result of [Bol01, Lemma 10.9] in studying the diameter of random graphs.

2.3 An Improved Algorithm in Sparse Graph Regime

In the sparse regime where np is poly-logarithmic, Algorithm 2 does not perform well. This is because for two distinct vertices u, v that are close by, their ℓ-th local neighborhoods have a large overlap, i.e., $|N_\ell^G(u) \cap N_\ell^G(v)|$ is not much smaller than $|N_\ell^G(u)|$ or $|N_\ell^G(v)|$, rendering w_{i_1,i_2} given in (15) ineffective to distinguish u from v.

However, in the sparse regime, distinct vertices u, v only have very few common neighbors. Moreover, if we remove vertices u, v, the non-common neighbors become far apart, and for distinct vertices far apart, their local neighborhoods only have a small overlap. Therefore, we expect most of u, v's neighbor's ℓ-th local neighborhoods (after removing vertices u, v) do not have large intersections for a suitably chosen ℓ. This gives rise to Algorithm 3.

Algorithm 3 Graph matching based on $(d-1)$-hop witness in sparse regime

1: **Input:** G_1, G_2, π_0, $\ell \in \mathbb{Z}$, $\eta \in \mathbb{R}_+$.
2: **Output:** $\widehat{\pi}$.
3: **Match high-degree vertices:** For all the pairs of unseeded vertices (u, v) and for each pair of their neighbors (i, j) with $i \in \Gamma_1^{G_1}(u)$ and $j \in \Gamma_1^{G_2}(v)$, compute

$$w_{i,j}^{u,v} = \min_{x \in V(G_1), y \in V(G_2)} \left\{ \left| \left\{ k \in I_0 : \pi_0(k) \in N_\ell^{G_1 \setminus \{u,x\}}(i),\ k \in N_\ell^{G_2 \setminus \{v,y\}}(j) \right\} \right| \right\}, \tag{16}$$

where $G \setminus S$ denotes G with set of vertices S removed. Let

$$Z_{u,v} = \sum_{i \in \Gamma_1^{G_1}(u)} \sum_{j \in \Gamma_1^{G_2}(v)} \mathbf{1}_{\{w_{i,j}^{u,v} \geq \eta\}}. \tag{17}$$

If $Z_{u,v} \geq \log n / \log \log n - 1$, set $\widehat{\pi}(v) = u$. Declare failure if there is any conflict.
4: The remaining two steps are the same as Algorithm 1.

Note that in computing the number of seeded vertices within distance ℓ from both vertex i in G_1 and vertex j in G_2 in (16), we remove vertices u, x in G_1 and vertices v, y in G_2, and take the minimum over all possible choices of x and y. As a result,

$$w_{i,j}^{u,v} \leq \left| \left\{ k \in I_0 : \pi_0(k) \in N_\ell^{G_1 \setminus \{u,v\}}(i),\ k \in N_\ell^{G_2 \setminus \{u,v\}}(j) \right\} \right|, \tag{18}$$

where the right hand side becomes independent from the edges incident to u and v in $G_1^* \vee G_2$. This independence is crucial in our analysis to ensure that $Z_{u,v}$ is small for $u \neq \pi^*(v)$ via concentration inequalities of multivariate polynomials [Vu02].

There are two tuning parameters ℓ and η in Algorithm 3. In our analysis later, we will optimally choose

$$\ell = \left\lfloor \frac{(1-\epsilon)\log n}{\log(np)} \right\rfloor, \tag{19}$$

and

$$\eta = 4^{2\ell+2} n^{1-2\epsilon} \alpha. \tag{20}$$

As for time complexity, Algorithm 3 takes at most $O(n^{5+2\epsilon})$ time steps. To see this, similar to Algorithm 2, if we assume one unit time to fetch a set of direct neighbors of a given vertex, then it takes at most $O(n^3)$ time steps to compute (18) for given pairs of vertices (u, v) and (i, j). There are at most $n^{2+2\epsilon}$ such pairs. The step of matching low-degree vertices as specified in Algorithm 1 takes $O(n^3 p)$ time steps in total. Thus in total Algorithm 3 takes at most $O(n^{5+2\epsilon})$ time steps.

2.4 Graph Matching without Seeds

Even without an initial seed set revealed as side information, we can select a random subset of vertices I_0 in G_1 and enumerate all the possible mappings $f : I_0 \to [n]$ from I_0 to vertices in G_2 in at most $n^{|I_0|}$ steps. Each of the possible mappings can be viewed as seeds; thus we can apply our seeded graph matching algorithm. Among all possible $n^{|I_0|}$ mappings, we finally output the best matching which minimizes the edge disagreements. See Algorithm 4 for details.

Algorithm 4 Seedless Graph matching via Seeded Graph Matching

1: **Input:** G_1, G_2
2: **Output:** $\widehat{\pi}$.
3: Select a random subset I_0 of $V(G_1)$ by including each vertex with probability α.
4: For every possible mapping $f : I_0 \to [n]$, run Seeded Graph Matching Algorithm with a seed set I_0, and output π_f.
5: Output

$$\widehat{\pi} \in \arg\min_{\pi_f} \|G_1 - \Pi_f G_2 \Pi^\top\|_F^2,$$

where Π_f is the permutation matrix corresponding to π_f.

Since one of the possible mapping f will correspond to the underlying true matches of vertices in I_0, it follows that if our seeded graph matching succeeds with high probability and we are above the information-theoretic limit (so that the true matching minimizes the edge disagreements with high probability), the final output will coincide with the true matching with high probability, as stated in Theorem 5. More specifically, the proof is sketched below.

Proof of Theorem 5. If $f : I_0 \to [n]$ is such that $f(i) = \pi^*(i)$ for all $i \in I_0$, then since our seeded graph matching succeeds with high probability, it follows that $\pi_f = \pi^*$ with high probability.

Moreover, since we are above the information-theoretic limit, it follows from [CK17, Theorem 1] that with high probability,

$$\pi^* \in \arg\min_{\pi} \|G_1 - \Pi G_2 \Pi^\top\|_F^2,$$

where Π is the permutation matrix corresponding to π.

Therefore, $\widehat{\pi} = \pi^*$ with high probability. Finally, since $\alpha n \to \infty$, it follows that $|I_0|$ is at most $2\alpha n$ with high probability. Hence, Algorithm 4 calls the Seeded Graph Matching algorithm at most $n^{O(\alpha n)}$ times with high probability. □

3 Analysis of Algorithm 1 in Sparse Graph Regime

In this and next two sections, we give the analysis of our algorithms and prove our main theorems. For the sake of analysis, we assume $\pi^* = id$, i.e., $\pi^*(i) = i$ for all $i \in [n]$, without loss of generality.

Our analysis of Algorithm 1 uses the technique for analyzing small subgraph containment [JLR11]. Let T denote a starlike tree formed by m independent (vertex-disjoint expect the root vertex) paths of length ℓ from root vertex to m distinct leaves for $\ell, m \geq 1$. Note that T consists of $m\ell+1$ vertices and $m\ell$ edges (See Fig. 1 for an example of $m = 3$ and $\ell = 2$). Let $r(T)$ denote the root vertex of T and $L(T)$ denote the set of leaves of T. We say T is a subgraph of G, denoted by $T \subset G$, if $V(T) \subset V(G)$ and $E(T) \subset E(G)$. The key of our proof is to show that under certain conditions with high probability:

1. For every vertex i, there exists a copy of T rooted at i with all leaves seeded in the intersection graph $G_1^* \wedge G_2$;
2. There is no copy of $T_1 \cup T_2$ in the union graph $G_1^* \vee G_2$.

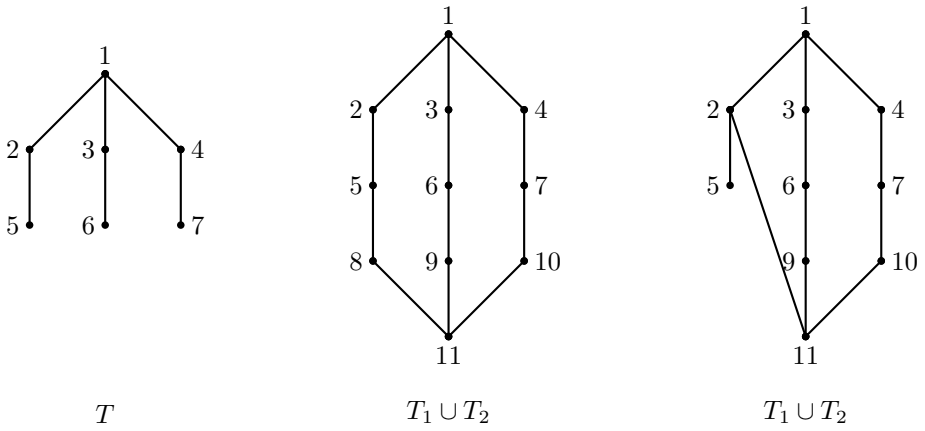

Figure 1: Left: T is a starlike tree with $m = 3$, $\ell = 2$, $r(T) = 1$ and $L(T) = \{5, 6, 7\}$. Middle and Right: Two examples of $T_1 \cup T_2$ such that T_1, T_2 are isomorphic to T, $r(T_1) \neq r(T_2)$, and $L(T_1) = L(T_2) = \{5, 6, 7\}$. For the middle, $V(T_1) \cap V(T_2) = \{5, 6, 7\}$; for the right, $V(T_1) \cap V(T_2) = \{2, 5, 6, 7\}$.

3.1 Success of Algorithm 1 on the Intersection of Good Events

We first introduce a sequence of "good" events on whose intersection, Algorithm 1 correctly matches all vertices. We need the following graph properties:

(i) there is no isolated vertex;

(ii) for any two adjacent vertices, there are at least τ vertices adjacent to at least one of them;

(iii) For all vertices i with $d_i \geq \tau$, there are at least $2m$ independent ℓ-paths from i to $2m$ distinct vertices in I_0;

(iv) There is no pairs of subgraphs $T_1, T_2 \subset G$ that are isomorphic to T such that $r(T_1) \neq r(T_2)$, and $L(T_1) = L(T_2)$ (See Fig. 1 for an illustration).

(v) For every vertex i, there exist at most $m-1$ independent ℓ-paths from i to $m-1$ distinct vertices in $N_{\ell-1}^{G}(i)$.

Let

- $\mathcal{E}_1$ denote the event such that $G_1^* \wedge G_2$ satisfy properties (i)–(iii);
- $\mathcal{E}_2$ denote the event such that $G_1^* \vee G_2$ satisfy properties (iv) and (v);
- $\mathcal{E}_3$ denote the event such that for any two vertices i, j that are connected by a 2-path in $G_1^* \vee G_2$, at least one of the two vertices i, j must be a high-degree vertex in $G_1^* \wedge G_2$.

We claim that on event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, Algorithm 1 correctly matches all vertices. Recall that we can assume $\pi^* = id$ and thus $G_1 = G_1^*$ without loss of generality.

First, since $G_1^* \wedge G_2$ satisfy graph property (iii), it follows that in $G_1^* \wedge G_2$, for all high-degree vertices i, there exist $2m$ independent ℓ-paths to a set $S \subset \Gamma_\ell^{G_1^* \wedge G_2}(i)$ of $2m$ seeded vertices. Let $\widetilde{S} = S \backslash N_{\ell-1}^{G_1^* \vee G_2}(i)$. Since $G_1^* \vee G_2$ satisfy graph property (v), and $G_1^* \wedge G_2 \subset G_1^* \vee G_2$, it follows that

$$\left| S \cap N_{\ell-1}^{G_1^* \vee G_2}(i) \right| \leq m-1$$

and thus $|\widetilde{S}| \geq m+1$. Moreover, since $G_1^* \wedge G_2 \subset G_1, G_2$, it follows that

$$\widetilde{S} \subset \Gamma_\ell^{G_1^* \wedge G_2}(i) \backslash N_{\ell-1}^{G_1^* \vee G_2} \subset \Gamma_\ell^{G_1^*}(i) \cap \Gamma_\ell^{G_2}(i).$$

Therefore, in both G_1 and G_2, there are at least $m+1$ independent ℓ-paths from i to $\Gamma_\ell^{G_1^*}(i) \cap \Gamma_\ell^{G_2}(i)$.

Second, note that on event $\mathcal{E}_2$, $G_1^* \vee G_2$ satisfy graph property (iv). For the sake of contradiction, suppose there exist a pair of distinct vertices i, j and a set L of m seeded vertices such that there exist m independent ℓ-paths from i to set L in G_1 and m independent ℓ-paths from j to set L in G_2. Let T_k denote the starlike tree formed by the m independent ℓ-paths in G_k for $k = 1, 2$. Then $T_1, T_2 \subset G_1^* \vee G_2$ are isomorphic to T such that $r(T_1) = i, r(T_2) = j$ and $L(T_1) = L(T_2) = L$. This is in contradiction with the fact that $G_1^* \vee G_2$ satisfy graph property (iv).

It follows from the above two points that Algorithm 1 correctly matches all high-degree vertices i in $G_1^* \wedge G_2$, i.e., $\widehat{\pi}(i) = i$.

Next, we show that all low-degree vertices are matched correctly. Fix a low-degree vertex i. Since $G_1^* \wedge G_2$ satisfy graph properties (i) and (ii), it must have a high-degree neighbor j in $G_1^* \wedge G_2$. Since the high-degree vertex j has been matched correctly, i is adjacent to j in G_1 and i is also adjacent to $\widehat{\pi}(j) = j$ in G_2. Moreover, for the sake of contradiction, suppose there exists a pair of two distinct low-degree vertices i_1 and i_2 such that i_1 is adjacent to a matched vertex j_1 in G_1 and i_2 is adjacent to vertex $\widehat{\pi}(j_1)$ in G_2. Since $\widehat{\pi}(j_1) = j_1$, it follows that (i_1, j_1, i_2) form a 2-path in $G_1^* \vee G_2$. However, on event $\mathcal{E}_3$, i_1 and i_2 cannot be low-degree vertices simultaneously in $G_1^* \wedge G_2$, which leads to a contradiction. As a consequence, $\widehat{\pi}(i) = i$ for every low-degree vertex i.

Finally, to prove Theorem 1, it remains to show that under the theorem assumptions, $\mathbb{P}\{\mathcal{E}_i\} \to 1$ for all $i = 1, 2, 3$, which is done in the next subsection.

3.2 Bound the Probability of Good Events

It is standard to prove that $G_1^* \wedge G_2$ satisfies properties (i)–(ii) with high probability and $\mathbb{P}\{\mathcal{E}_3\} \to 1$ using union bounds. For completeness, we state the lemmas and leave the proofs to appendices.

Lemma 1. *Suppose $G \sim \mathcal{G}(n,p)$ with $np - \log n \to +\infty$.*

(i) There is no isolated vertex in G with probability at least $1 - o(1)$;

(ii) Assume $\tau = o(np)$. With probability at least $1 - n^{-1+o(1)}$, for any two adjacent vertices, there are at least τ vertices adjacent to at least one of them in G.

Lemma 2. *Assume*

$$nps^2 \geq \log n \quad \text{and} \quad \tau = o(nps^2) \quad \text{and} \quad \log(np) = o(nps^2).$$

With probability at least $1 - n^{-1+o(1)}$, for any two vertices i, j that are connected by a 2-path in $G_1^* \vee G_2$, at least one of the two vertices i, j must have degree at least τ in $G_1^* \wedge G_2$.

It remains to show with high probability, $G_1^* \wedge G_2$ satisfy graph property (iii) and $G_1^* \vee G_2$ satisfy graph properties (iv) and (v).

We will apply the following lemma to show that with high probability, for every high-degree vertex i in $G_1^* \wedge G_2$, we can always find at least $2m$ independent paths of length ℓ from i to $2m$ distinct seeded vertices in I_0.

Lemma 3. *Suppose $G \sim \mathcal{G}(n,p)$ and each vertex in G is included in I_0 independently with probability α. Assume*

$$\alpha(np/2)^{\ell-2}\tau(\tau - 2m) - 2m\log\tau \geq 2\log n.$$

and

$$p(4np)^{\ell} = o(1)$$

and $\tau \to +\infty$. Then with high probability, for all vertices i with $d_i \geq \tau$, there are at least $2m$ independent ℓ-paths from i to I_0.

Proof. In view of Proposition 1, we have $\mathbb{P}\{\mathcal{H}\} \geq 1 - 3n^{-1+o(1)}$, where on event $\mathcal{H}$, for every vertex i, there exists a tree $T_\ell(i) \subset G$ of depth ℓ rooted at i such that:

1. Root i has at most one children j who has fewer than τ children in $T(i)$, i.e., $|\Pi_1(j)| \leq \tau$.
2. For each children j of i with $|\Pi_1(j)| \geq \tau$, the subtree $T_{\ell-1}(j)$ of depth $\ell - 1$ rooted at j has at least $\tau(np/2)^{\ell-2}$ leaves, i.e., $|\Pi_{\ell-1}(j)| \geq \tau\,(np/2)^{\ell-2}$.

Fix a vertex i in G. Then i has at least $d_i - 1$ children j such that $|\Pi_1(j)| \leq \tau$. For each such j, define $Y_{ij} = 1$ if there is a path of length $\ell - 1$ from j to some vertex in I_0 in $T_\ell(i)$. Then the number of independent paths from i to I_0 is at least $\sum_{j=1}^{d_i - 1} Y_{ij}$.

Since each leaf vertex of $T_{\ell-1}(j)$ is included in I_0 with probability α independently across different vertices and from graph G, it follows that

$$\mathbb{P}\{Y_{ij} = 1 \mid d_i \geq \tau, \mathcal{H}\} = 1 - (1-\alpha)^{|\Pi_{\ell-1}(j)|} \geq 1 - \exp\left(-\alpha\tau\,(np/2)^{\ell-2}\right),$$

where we used $1-x \le e^{-x}$ and $|\Pi_{\ell-1}(j)| \ge \tau\,(np/2)^{\ell-2}$ on event $\mathcal{H}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left\{\sum_{j=1}^{d_i-1} Y_{ij} \le 2m-1 | d_i \ge \tau, \mathcal{H}\right\} &\le \mathbb{P}\left\{\sum_{j=1}^{\tau-1} Y_{ij} \le 2m-1\right\} \\
&\le \mathbb{P}\left\{\mathrm{Binom}\left(\tau-1, 1-e^{-\alpha\tau(np/2)^{\ell-2}}\right) \le 2m-1\right\} \\
&= \sum_{k=0}^{2m-1} \binom{\tau-1}{k} e^{-\alpha\tau(np/2)^{\ell-2}(\tau-1-k)} \\
&\le e^{-\alpha\tau(np/2)^{\ell-2}(\tau-2m)} \sum_{k=0}^{m-1} \tau^k \\
&\le 2e^{-\alpha\tau(np/2)^{\ell-2}(\tau-2m)}\tau^{2m} \le 2n^{-2},
\end{aligned}$$

where the last equality holds due to the assumption $\alpha(\tau-2m)\,(np/2)^{\ell-2} - 2m\log\tau \ge 2\log n$.

Define event

$$\mathcal{F}_i = \{d_i \ge \tau\} \cap \left\{\sum_{j=1}^{d_i-1} Y_{ij} \le 2m-1\right\}.$$

Then we have that

$$\mathbb{P}\{\mathcal{F}_i \cap \mathcal{H}\} \le \mathbb{P}\left\{\sum_{j=1}^{d_i-1} Y_{ij} \le 2m-1 | d_i \ge \tau, \mathcal{H}\right\} \le 2n^{-2}.$$

Let $\mathcal{F} = \cup_i \mathcal{F}_i$. By the union bound,

$$\mathbb{P}\{\mathcal{F}\} = \mathbb{P}\{\mathcal{F}\cap\mathcal{H}\} + \mathbb{P}\{\mathcal{H}^c\} \le \sum_i \mathbb{P}\{\mathcal{F}_i\cap\mathcal{H}\} + \mathbb{P}\{\mathcal{H}^c\} \le 2n^{-1} + 3n^{-1+o(1)} \le 5n^{-1+o(1)}.$$

Therefore, with high probability, for all vertices i with $d_i \ge \tau$, $\sum_{j=1}^{d_i-1} Y_{ij} \ge 2m$.

□

The following lemma will be useful to conclude that in $G_1^* \vee G_2$, with high probability, there is no pair of subgraphs $T_1, T_2 \subset G_1^* \vee G_2$ that are isomorphic to T, such that $r(T_1) \ne r(T_2)$ and $L(T_1) = L(T_2)$. See Fig. 1 for an illustration of T_1 and T_2 isomorphic to T such that $r(T_1) \ne r(T_2)$ and $L(T_1) = L(T_2)$.

Lemma 4. *Suppose $G \sim \mathcal{G}(n,p)$ and $\ell, m \ge 1$. Then it holds that*

$$\begin{aligned}
&\mathbb{P}\{\exists\, T_1, T_2 \subset G \text{ that are isomorphic to } T : r(T_1) \ne r(T_2), L(T_1) = L(T_2)\} \\
&\le \left(2 + \frac{8}{np}\right)^{m(\ell-1)} n^{2m\ell+2-m} p^{2m\ell}.
\end{aligned}$$

Proof. Let $\mathcal{T}$ denote the set of all possible subgraphs that are isomorphic to T in the complete graph K_n. By the union bound, we have

$$\begin{aligned}
&\mathbb{P}\{\exists\, T_1, T_2 \subset G \text{ that are isomorphic to } T : r(T_1) \ne r(T_2), L(T_1) = L(T_2)\} \\
&\le \sum_{T_1, T_2 \in \mathcal{T}:\, r(T_1)\ne r(T_2), L(T_1)=L(T_2)} \mathbb{P}\{T_1, T_2 \subset G\}
\end{aligned}$$

For each such pair of T_1, T_2,

$$\mathbb{P}\{T_1, T_2 \subset G\} = p^{|E(T_1)|+|E(T_2)|-|E(T_1\cap T_2)|} = p^{2m\ell-|E(T_1\cap T_2)|},$$

where the last equality holds because T_1 and T_2 are isomorphic to T and $|E(T)| = 2m\ell$.

Next for any given unlabelled graph S, we enumerate all the possible distinct pairs of $T_1, T_2 \in \mathcal{T}$ such that $T_1 \cap T_2$ is isomorphic to S, $r(T_1) \neq r(T_2)$, and $L(T_1) = L(T_2)$. Let κ_S denote the number of subgraphs S' in T such that S' is isomorphic to S, $L(T) \subset V(S')$, and $r(T) \notin V(S')$. Then there are at most κ_S^2 ways of intersecting T_1 and T_2 such that $T_1 \cap T_2$ is isomorphic to S, $r(T_1) \neq r(T_2)$, and $L(T_1) = L(T_2)$. For each such type of intersection, there are at most $n^{|V(S)|}$ different choices for vertex labelings of $T_1 \cap T_2$, and $n^{2(|V(T)|-|V(S)|)}$ different choices for vertex labelings of $(T_1 \backslash T_2) \cup (T_2 \backslash T_1)$. Hence, the total number of distinct pairs of $T_1, T_2 \in \mathcal{T}$ such that $T_1 \cap T_2$ is isomorphic to S, $r(T_1) \neq r(T_2)$, and $L(T_1) = L(T_2)$ is at most

$$\kappa_S^2 n^{|V(S)|} n^{2(|V(T)|-|V(S)|)} = \kappa_S^2 n^{2m\ell+2-|V(S)|},$$

where the last equality holds due to $|V(T)| = m\ell + 1$.

Combining the last two displayed equations yields that

$$\sum_{T_1,T_2\in\mathcal{T}:\, r(T_1)\neq r(T_2), L(T_1)=L(T_2)} \mathbb{P}\{T_1, T_2 \subset G\} \leq \sum_S \kappa_S^2 n^{2m\ell+2-|V(S)|} p^{2m\ell-|E(S)|}.$$

Note that if $\kappa_S \geq 1$, then by the definition of κ_S, S is isomorphic to some $S' \subset T$ such that $L(T) \subset V(S')$ and $r(T) \notin V(S')$. By the starlike tree property of T, S' is a forest with at least m disjoint trees; hence so is S. See Fig. 1 for two illustrating examples. Therefore,

$$E(S) \leq V(S) - m.$$

Hence,

$$\sum_S \kappa_S^2 n^{2m\ell+2-|V(S)|} p^{2m\ell-|E(S)|} \leq \sum_S \kappa_S^2 n^{2m\ell+2-|V(S)|} p^{2m\ell+m-|V(S)|}.$$

Finally, we break the summation in the right hand side of the last displayed equation according to $|V(S)|$. In particular, let $|V(S)| = m + k$ for $0 \leq k \leq m(\ell-1)$. Note that $\sum_S \kappa_S$ is at most the number of distinct subgraphs S' of T such that $L(T) \subset V(S'), r(T) \notin V(S')$ and $|V(S')| = m + k$, which is further upper bounded by $\binom{m(\ell-1)}{k} 2^k$, because there are at most $\binom{m(\ell-1)}{k}$ different choices for $V(S') \backslash L(T)$ and at most $2^{|V(S')|-m}$ choices for determining whether to include the edges induced by $V(S')$ in T into S'. Hence,

$$\begin{aligned}
&\sum_S \kappa_S^2 n^{2m\ell+2-|V(S)|} p^{2m\ell+m-|V(S)|} \\
&= \sum_{k=0}^{m(\ell-1)} n^{2m\ell+2-m-k} p^{2m\ell-k} \sum_{S:|V(S)|=m+k} \kappa_S^2 \\
&\overset{(a)}{\leq} \sum_{k=0}^{m(\ell-1)} n^{2m\ell+2-m-k} p^{2m\ell-k} \left(\binom{m(\ell-1)}{k} 2^k \right)^2 \\
&\overset{(b)}{\leq} n^{2m\ell+2-m} p^{2m\ell} 2^{m(\ell-1)} \sum_{k=0}^{m(\ell-1)} n^{-k} p^{-k} \binom{m(\ell-1)}{k} 4^k \\
&= n^{2m\ell+2-m} p^{2m\ell} 2^{m(\ell-1)} \left(1 + \frac{4}{np}\right)^{m(\ell-1)},
\end{aligned}$$

where (a) follows from $\sum_S \kappa_S \le \binom{m(\ell-1)}{k} 2^k$, and (b) holds due to $\binom{m(\ell-1)}{k} \le 2^{m(\ell-1)}$. $\square$

Finally, we need a result to conclude that with high probability, for every vertex i, there exist at most $m-1$ independent ℓ-paths from i to $m-1$ distinct vertices in $N_{\ell-1}^{G_1^* \vee G_2}(i)$.

Fix $m, \ell \ge 1$. We start with any vertex i and m independent (vertex-disjoint except for i) paths of length ℓ from i to m distinct vertices $j_1, \ldots, j_m$, denoted by $P_1, \ldots, P_m$. Let $\widetilde{P}_k$ denote any path of length at most $\ell-1$ from i to j_k for $k = 1, \ldots, m$. Let $H = \cup_{k=1}^m (P_k \cup \widetilde{P}_k)$ and $\mathcal{H}_{m,\ell}$ denote the family of all possible graphs H with $V(H) \subset [n]$ obtained by the above procedure.

Note that if there is no subgraph isomorphic to some $H \in \mathcal{H}_{m,\ell}$ in $G_1^* \vee G_2$, then for every vertex i, there exist at most $m-1$ independent ℓ-paths from i to $m-1$ distinct vertices in $N_{\ell-1}^{G_1^* \vee G_2}(i)$. Hence, our task reduces to proving that with high probability, $G_1^* \vee G_2$ does not contain some $H \in \mathcal{H}_{m,\ell}$ as a subgraph.

We first need a lemma showing that any $H \in \mathcal{H}_{m,\ell}$ is so "dense" that it appears as a subgraph in $\mathcal{G}(n,p)$ with a vanishing small probability.

Lemma 5. *Fix $m, \ell \ge 1$. For any $H \in \mathcal{H}_{m,\ell}$,*

$$|E(H)| \ge |V(H)| + m - 1.$$

Proof. Recall that $H = \cup_{k=1}^m (P_k \cup \widetilde{P}_k)$, where $P_1, \ldots, P_m$ is a set of m (vertex-disjoint except for i) paths of length ℓ from i to m distinct vertices $j_1, \ldots, j_m$, and $\widetilde{P}_k$ is a path of length at most $\ell-1$ from i to j_k for $k = 1, \ldots, m$.

Note that we order the vertices and edges in paths starting from i. For each $k = 1, \ldots, m$, let v_k denote the first vertex after which P_k and $\widetilde{P}_k$ completely conincide, and e_k denote the edge incident to v_k in $\widetilde{P}_k$. Then by definition, $v_k \ne i$ and $e_k \in \widetilde{P}_k \backslash P_k$. Let $\mathrm{dist}(u,v)$ denote the *longest* distance between u and v in H, and σ denote any permutation on $[m]$ such that

$$\mathrm{dist}\left(i, v_{\sigma(1)}\right) \ge \mathrm{dist}\left(i, v_{\sigma(2)}\right) \cdots \ge \mathrm{dist}\left(i, v_{\sigma(m)}\right).$$

Without loss of generality, we assume $\sigma = id$, i.e., $\sigma(k) = k$. We claim that $e_j \notin P_k \cup \widetilde{P_k}$ for any $1 \le j < k \le m$. In fact, $e_j \notin P_k$, because otherwise P_j and P_k share a common vertex $v_j \ne i$, which violates the assumption that P_j and P_k are vertex-disjoint except for i. Also, $e_j \notin \widetilde{P}_k \backslash P_k$, because otherwise, e_j is ordered before e_k in path $\widetilde{P}_k$ starting from i, which implies $\mathrm{dist}(i, v_k) > \mathrm{dist}(i, v_j)$ and leads to a contradiction.

Finally, we recursively define $H_0 = H$ and H_k such that $V(H_k) = V(H_{k-1})$ and $E(H_k) = E(H_{k-1}) \backslash \{e_k\}$ for $k = 1, \ldots, m$. We prove that H_m is connected by induction. For the base case $k = 0$, clearly $H_0 = H$ is connected. Suppose H_{k-1} is connected. Since we have shown that $e_j \notin P_k \cup \widetilde{P_k}$ for any $1 \le j < k \le m$, it follows that $P_k \cup \widetilde{P_k} \subset H_{k-1}$. Note that there is a path through i between the two endpoints of e_k in $P_k \cup \widetilde{P_k}$. Hence, the two endpoints of e_k are still connected in H_k. Moreover, by the induction hypothesis, H_{k-1} is connected. Therefore, H_k is connected. and it follows from induction that H_m is connected. Thus, $|E(H_m)| - |V(H_m)| \ge -1$. Since $|E(H)| = |E(H_m)| + m$ and $|V(H)| = |V(H_m)|$, it follows that $|E(H)| - |V(H)| \ge m - 1$. $\square$

Next we state a lemma which upper bounds the number of isomorphism classes in $\mathcal{H}_{m,\ell}$. This upper bound is by no means tight, but suffices for our purpose.

Lemma 6. *Fix $m, \ell \ge 1$. Denote by $\mathcal{U}_{m,\ell}$ the set of unlabelled graphs (isomorphism classes) in $\mathcal{H}_{m,\ell}$. Then*

$$|\mathcal{U}_{m,\ell}| \le (3\ell)^m 3^{2m^2\ell}.$$

Proof. Recall that $H = \cup_{k=1}^{m}(P_k \cup \widetilde{P}_k)$, where $P_1, \ldots, P_m$ is a set of m (vertex-disjoint except for i) paths of length ℓ from i to m distinct vertices $j_1, \ldots, j_m$, and $\widetilde{P}_k$ is a path of length at most $\ell - 1$ from i to j_k for $k = 1, \ldots, m$. Let $T = \cup_{k=1}^{m} P_k$. Then T is a starlike tree rooted at i with m branches as depicted in Fig. 1.

We fix a sequence of $\{\ell_1, \ldots, \ell_m\}$ with $1 \le \ell_k \le \ell - 1$. Let $\mathcal{U}_{\ell_1,\ldots,\ell_m}$ denote all the possible unlabelled graphs formed by the union of T and $\widetilde{P}_k$ of length ℓ_k for $k \in [m]$. For ease of notation, let $\widetilde{P}_0 = T$. We enumerate $\mathcal{P}_{\ell_1,\ldots,\ell_m}$ according to the pairwise intersections $\widetilde{P}_j \cap \widetilde{P}_k$ for $0 \le j < k \le m$. Specifically, for any given sequence $\{S_{jk} : 0 \le j < k \le m\}$ of unlabelled graphs, we enumerate all the possible sequences of $(\widetilde{P}_1, \ldots, \widetilde{P}_k)$ such that $\widetilde{P}_j \cap \widetilde{P}_k$ is isomorphic to S_{jk} for $0 \le j < k \le m$. Let $\kappa_\ell(S)$ denote the number of possible different subgraphs that are isomorphic to S in an ℓ-path. Recall $\beta(S)$ denote the number of possible different subgraphs that are isomorphic to S in $\widetilde{P}_0 = T$.

Then across all $1 \le j < k \in [m]$, there are at most $\kappa_{\ell_j}(S)\kappa_{\ell_k}(S)$ ways of intersecting $\widetilde{P}_j$ and $\widetilde{P}_k$ such that $\widetilde{P}_j \cap \widetilde{P}_k$ is isomorphic to S. Also, for all $k \in [m]$, there are at most $\beta(S)\kappa_{\ell_k}(S)$ ways of intersecting $\widetilde{P}_0$ and $\widetilde{P}_k$ such that $\widetilde{P}_0 \cap \widetilde{P}_k$ is isomorphic to S. Hence, the total number of distinct sequences of $(\widetilde{P}_1, \ldots, \widetilde{P}_k)$ such that $\widetilde{P}_j \cap \widetilde{P}_k$ isat most the the number n_ℓ of distinct subgraphs in an ℓ-path isomorphic to S_{jk} for $0 \le j < k \le m$ is at most

$$\prod_{1\le j<k\in[m]} \kappa_{\ell_j}(S_{jk})\kappa_{\ell_k}(S_{jk}) \prod_{k\in[m]} \beta(S_{0k})\kappa_{\ell_k}(S_{0k}).$$

Therefore,

$$\begin{aligned}
|\mathcal{U}_{\ell_1,\ldots,\ell_m}| &\le \sum_{\{S_{jk}:0\le j<k\le m\}} \prod_{j<k\in[m]} \kappa_{\ell_j}(S_{jk})\kappa_{\ell_k}(S_{jk}) \prod_{k\in[m]} \beta(S_{0k})\kappa_{\ell_k}(S_{0k}) \\
&\le \prod_{1\le j<k\in[m]} \left(\sum_{S_{jk}} \kappa_{\ell_j}(S_{jk})\right)\left(\sum_{S_{jk}} \kappa_{\ell_k}(S_{jk})\right) \prod_{k\in[m]} \left(\sum_{S_{0k}} \beta(S_{0k})\right)\left(\sum_{S_{0k}} \kappa_{\ell_k}(S_{0k})\right) \\
&\le \prod_{j<k\in[m]} n_{\ell_j} n_{\ell_k} \prod_{k\in[m]} n(T) n_{\ell_k} = (n(T))^m \prod_{k\in[m]} (n_{\ell_k})^m,
\end{aligned}$$

where the last inequality holds because $\sum_S \kappa_\ell(S)$ is at most the the number n_ℓ of distinct subgraphs S' in an ℓ-path, and $\sum_S \beta(S)$ is at most the the number $n(T)$ of distinct subgraphs S' in T. Note that

$$n_\ell \le \sum_{k=0}^{\ell} \binom{\ell}{k} 2^k = 3^\ell,$$

because if $|V(S')| = k$, then there are at most $\binom{\ell}{k}$ different choices for $V(S')$ and at most 2^k choices for determining whether to include the edges induced by $V(S')$ in an ℓ-path into S'. Also,

$$n(T) \le \sum_{k=0}^{m\ell+1} \binom{m\ell+1}{k} 2^k = 3^{m\ell+1}.$$

Combinining the last three displayed equations yields that

$$|\mathcal{U}_{\ell_1,\ldots,\ell_m}| \le 3^{2m^2\ell+m}.$$

Therefore,

$$|\mathcal{U}| = \sum_{(\ell_1,\ldots,\ell_m):1\le\ell_k\le\ell-1} |\mathcal{U}_{\ell_1,\ldots,\ell_m}| \le (3\ell)^m 3^{2m^2\ell}.$$

□

With Lemma 5 and Lemma 6, we are ready to bound the probability that $\mathcal{G}(n,p)$ contains some $H \in \mathcal{H}_{m,\ell}$ as a subgraph.

Lemma 7. *Suppose $G \sim \mathcal{G}(n,p)$ with $np \geq 1$ and $m, \ell \geq 1$. Then it holds that*

$$\mathbb{P}\left\{\exists H \in \mathcal{H}_{m,\ell} : H \subset G\right\} \leq n^{2m\ell-2m+1} p^{2m\ell-m} (3\ell)^m 3^{2m^2\ell}. \tag{21}$$

Proof. Note that for any $H \in \mathcal{H}_{m,\ell}$,

$$m\ell + 1 \leq |V(H)| \leq m\ell + 1 + (\ell-2)m = 2m\ell - 2m + 1,$$

where the lower bound holds because H contains a starlike tree with $m\ell+1$ distinct vertices, and the upper bound holds when P_k and $\widetilde{P}_k$ are all vertex-disjoint except for the source vertex and sink vertices.

For any given integer $m\ell + 1 \leq t \leq 2m\ell - 2m + 1$, define

$$\mathcal{H}_{m,\ell,t} = \left\{H \in \mathcal{H}_{m,\ell} : V(H) \subset [n], |V(H)| = t\right\}.$$

and let $\mathcal{U}_{m,\ell_t}$ denote the number of unlabelled graphs (isomorphism class) in $\mathcal{H}_{m,\ell}$. Since $V(H) \subset [n]$ and $|V(H)| = t$, there are at most n^t different vertex labelings for a given unlabelled graph $U \in \mathcal{U}_{m,\ell,t}$. Hence,

$$|\mathcal{H}_{m,\ell,t}| \leq |\mathcal{U}_{m,\ell,t}|\, n^t. \tag{22}$$

By the union bound, we have

$$\begin{aligned}
\mathbb{P}\left\{\exists H \in \mathcal{H}_{m,\ell} : H \subset G\right\} &\leq \sum_{t=m\ell+1}^{2m\ell-2m+1} \sum_{H\in\mathcal{H}_{m,\ell,t}} \mathbb{P}\left\{H \subset G\right\} \\
&= \sum_{t=m\ell+1}^{2m\ell-2m+1} \sum_{H\in\mathcal{H}_{m,\ell,t}} p^{|E(H)|} \\
&\overset{(a)}{\leq} \sum_{t=m\ell+1}^{2m\ell-2m+1} \sum_{H\in\mathcal{H}_{m,\ell,t}} p^{t+m-1} \\
&\overset{(b)}{\leq} \sum_{t=m\ell+1}^{2m\ell-2m+1} n^t p^{t+m-1} |\mathcal{U}_{m,\ell,t}| \\
&\overset{(c)}{\leq} n^{2m\ell-2m+1} p^{2m\ell-m} |\mathcal{U}_{m,\ell}| \\
&\leq n^{2m\ell-2m+1} p^{2m\ell-m} (3\ell)^m 3^{2m^2\ell},
\end{aligned}$$

where (a) holds in view of Lemma 5; (b) holds in view of (22) (c) holds because $np \geq 1$ and $t \leq 2m\ell - 2m + 1$; the last inequality holds due to Lemma 6. □

3.3 Completing the Proof of Theorem 1

Recall that the choices of ℓ in (12), m in (13), and τ in (14). In particular,

$$\left(nps^2\right)^\ell \leq n^{1/2-\epsilon}.$$

Recall that $G_1^* \wedge G_2 \sim \mathcal{G}(n, ps^2)$. Under the assumption that $\alpha \geq n^{-1/2+3\epsilon}$, we get that

$$\alpha(nps^2/2)^{\ell-2}\tau(\tau - m) - m\log\tau \geq 2\log n.$$

Hence, applying Lemma 3, we conclude that $G_1^* \wedge G_2$ satisfy graph property (iii). Combing this result with Lemma 1, we get that $\mathbb{P}\{\mathcal{E}_1\} \geq 1 - o(1)$.

Note that $G_1^* \vee G_2 \sim \mathcal{G}(n, ps(2-s))$. We first apply Lemma 4 to $G_1^* \vee G_2$. In view of $nps^2 \geq \log n$ and $n \geq e$, we get that $nps(2-s) \geq \log n \geq 1$ and thus

$$\begin{aligned}
&\left(2 + \frac{8}{nps(2-s)}\right)^{m(\ell-1)} n^{2m\ell+2-m} \left(ps(2-s)\right)^{2m\ell} \\
&\leq 10^{m\ell} n^{2-2\epsilon m} \left(\frac{2-s}{s}\right)^{2ml} = n^{-2+o(1)},
\end{aligned}$$

where the first inequality holds due to $(nps^2)^\ell \leq n^{1/2-\epsilon}$; the last equality holds by our choice of ℓ and m and $s = \Theta(1)$. Hence, applying Lemma 4 to $G_1^* \vee G_2$, we conclude that with high probability, there is no pair of subgraphs $T_1, T_2 \subset G_1^* \vee G_2$ that are isomorphic to T such that $r(T_1) \neq r(T_2)$ and $L(T_1) = L(T_2)$.

Then we apply Lemma 7 to $G_1^* \vee G_2$. Note that

$$\begin{aligned}
&n^{2m\ell-2m+1} \left(ps(2-s)\right)^{2m\ell-m} (3\ell)^m 3^{2m^2\ell} \\
&\leq n^{m(1-2\epsilon)-m+1} (np)^{-m} \left(\frac{2-s}{s}\right)^{2m\ell-m} (3\ell)^m 3^{2m^2\ell} \leq n^{-3+o(1)},
\end{aligned}$$

where the first inequality holds due to $(nps^2)^\ell \leq n^{1/2-\epsilon}$; the last equality holds by our choice of ℓ and m and $s = \Theta(1)$. Hence, applying Lemma 7 to $G_1^* \vee G_2$, we conclude that with high probability, $G_1^* \vee G_2$ does not contain any graph $H \in \mathcal{H}_{m,\ell}$ as a subgraph. By the construction of $\mathcal{H}_{m,\ell}$, it further implies that with high probability, for every vertex i, there exist at most $m-1$ independent paths from i to $m-1$ distinct vertices in $N_{\ell-1}^{G_1^* \vee G_2}$.

Combining the above two points, we get that $\mathbb{P}\{\mathcal{E}_2\} \to 1$. Finally, in view of Lemma 2, we get that $\mathbb{P}\{\mathcal{E}_3\} \geq 1 - o(1)$, completing the proof of Theorem 1.

4 Analysis of Algorithm 2 in Dense Graph Regime

Recall that $\Gamma_G^k(u)$ and $N_k^G(u)$ denotes the set of vertices *at* and *within* distance k from u in graph G, respectively, as defined in (10) and (11). The key is to show that $|N_{d-1}^{G_1^* \wedge G_2}(u)|$ is larger than $|N_{d-1}^{G_1^* \vee G_2}(u) \cap N_{d-1}^{G_1^* \vee G_2}(v)|$ for $u \neq v$ by a constant factor, so that we can matches two vertices correctly based on the number of common seeded vertices in their two large neighborhoods.

Proof of Theorem 4. Define event

$$\mathcal{A} = \left\{ \left| N_{d-1}^{G_1^* \wedge G_2}(u) \right| \geq \frac{3}{4}(nps^2)^{d-1},\ \forall u \right\}.$$

In view of claim (i) in Lemma 11 with $G = G_1^* \wedge G_2$ and the fact that $\Gamma_k^G(u) \subset N_k^G(u)$, we get that $\mathbb{P}\{\mathcal{A}\} \geq 1 - n^{-10}$.

Define event

$$\mathcal{B} = \left\{ \left| N_{d-1}^{G_1^* \vee G_2}(u) \cap N_{d-1}^{G_1^* \vee G_2}(v) \right| \leq \frac{1}{2}(nps^2)^{d-1},\ \forall u \neq v \right\}.$$

Note that due to assumption (8),

$$\frac{1}{2}(nps^2)^{d-1} \geq 8n^{2d-3}\left(ps(2-s)\right)^{2d-2}.$$

Hence, applying claim (ii) in Lemma 11 with $G = G_1^* \vee G_2$, we get that $\mathbb{P}\{\mathcal{B}\} \geq 1 - n^{-10}$.

Recall that I_0 is the initial set of seeded vertices. Define event

$$\mathcal{C} = \left\{ \left| N_{d-1}^{G_1^* \wedge G_2}(u) \cap I_0 \right| > \frac{3}{5}(nps^2)^{d-1}\alpha,\ \forall u \right\}.$$

Since each vertex is seeded independently with probability α, it follows that

$$\begin{aligned}
\mathbb{P}\{\mathcal{C}^c\} &\leq \mathbb{P}\{\mathcal{A}^c\} + \mathbb{P}\{\mathcal{C}^c \mid \mathcal{A}\} \\
&\leq n^{-10} + \sum_u \mathbb{P}\left\{ \left| N_{d-1}^{G_1^* \wedge G_2}(u) \cap I_0 \right| \leq \frac{3}{5}(nps^2)^{d-1}\alpha \mid \mathcal{A} \right\} \\
&\leq n^{-10} + n\mathbb{P}\left\{ \mathrm{Bin}\left(\left\lceil \frac{3}{4}(nps^2)^{d-1} \right\rceil, \alpha \right) \leq \frac{3}{5}(nps^2)^{d-1}\alpha \right\} \\
&\leq n^{-10} + n\exp\left(-\frac{3}{200}(nps^2)^{d-1}\alpha \right) \leq 2n^{-1},
\end{aligned}$$

where the last inequality holds due to assumption (9).

Similarly, define event

$$\mathcal{D} = \left\{ \left| N_{d-1}^{G_1^* \vee G_2}(u) \cap N_{d-1}^{G_1^* \vee G_2}(v) \cap I_0 \right| < \frac{3}{5}(nps^2)^{d-1}\alpha,\ \forall u \neq v \right\}.$$

It follows that

$$\begin{aligned}
\mathbb{P}\{\mathcal{D}^c\} &\leq \mathbb{P}\{\mathcal{B}^c\} + \mathbb{P}\{\mathcal{D}^c \mid \mathcal{B}\} \\
&\leq n^{-10} + \sum_u \mathbb{P}\left\{ \left| N_{d-1}^{G_1^* \vee G_2}(u) \cap N_{d-1}^{G_1^* \vee G_2}(v) \cap I_0 \right| \geq \frac{3}{5}(nps^2)^{d-1}\alpha \mid \mathcal{B} \right\} \\
&\leq n^{-10} + n\mathbb{P}\left\{ \mathrm{Bin}\left(\left\lceil \frac{1}{2}(nps^2)^{d-1} \right\rceil, \alpha \right) \leq \frac{3}{5}(nps^2)^{d-1}\alpha \right\} \\
&\leq n^{-10} + n\exp\left(-\frac{1}{150}(nps^2)^{d-1}\alpha \right) \leq 2n^{-1},
\end{aligned}$$

where the last inequality holds again due to assumption (9). Hence, $\mathbb{P}\{\mathcal{C} \cap \mathcal{D}\} \geq 1 - 4n^{-1}$.

Finally, since $G_1^* \wedge G_2$ is a subgraph of both G_1^* and G_2, it follows that

$$N_{d-1}^{G_1^* \wedge G_2}(i_2) \subset \left\{ j \in I_0 : \pi_0(j) \in N_{d-1}^{G_1}\left(\pi^*(i_2)\right), j \in N_{d-1}^{G_2}(i_2) \right\}.$$

Similarly, both G_1^* and G_2 are subgraphs of $G_1^* \vee G_2$, it follows that

$$\left\{ j \in I_0 : \pi_0(j) \in N_{d-1}^{G_1}\left(i_1\right), j \in N_{d-1}^{G_2}(i_2) \right\} \subset N_{d-1}^{G_1^* \vee G_2}\left((\pi^*)^{-1}(i_1)\right) \cap N_{d-1}^{G_1^* \vee G_2}(i_2).$$

Thus, on event $\mathcal{C} \cap \mathcal{D}$, for every vertex $i_2 \in V(G_2) \setminus I_0$,

$$w_{i_1,i_2} \begin{cases} > \frac{3}{5}(nps^2)^{d-1}\alpha & \text{if } i_1 = \pi^*(i_2) \\ < \frac{3}{5}(nps^2)^{d-1}\alpha & \text{o.w. .} \end{cases}$$

Hence, Algorithm 2 outputs $\widehat{\pi} = \pi^*$ on event $\mathcal{C} \cap \mathcal{D}$. □

5 Analysis of Algorithm 3 in Sparse Graph Regime

Recall that we assume $\pi^* = id$ without loss of generality in the analysis. Before proving Theorem 3, we present two key lemmas.

The first lemma will be used later to conclude that the test statistic $Z_{u,u}$ given in (17) is large for all high degree vertices u.

Lemma 8. *Suppose $G \sim \mathcal{G}(n,p)$ with $\log n \le np \le n^{\epsilon}$, and each vertex is included in I_0 with probability α. Recall that ℓ and η are given in (19) and (20), respectively. Assume $\eta \ge 4\log n$. Let $G\backslash S$ denote the graph G with set of vertices S removed. Then with probability at least $1-n^{-1+o(1)}$,*

$$\sum_{j\in\Gamma_1^G(i)} \mathbf{1}_{\left\{|\Gamma_\ell^{G\backslash S}(j)\cap I_0|\ge\eta\right\}} \ge d_i - 1, \quad \forall S \text{ s.t. } i\in S, |S|\le 3.$$

Proof. For every vertex i and its neighbor $j\in\Gamma_1^G(i)$, define

$$a_{ij} = \mathbf{1}_{\left\{|\Gamma_\ell^{G\backslash S}(j)|\ge 4\eta/\alpha\right\}}.$$

and

$$b_{ij} = \mathbf{1}_{\left\{|\Gamma_\ell^{G\backslash S}(j)\cap I_0|\ge\eta\right\}}.$$

Define event

$$\mathcal{A} = \left\{|\Gamma_\ell^{G\backslash S}(j)| \ge \frac{(np)^\ell}{2^{\ell-1}\log(np)},\ \forall S \text{ s.t. } |S|\le 3,\ \forall j \text{ s.t. } \frac{np}{\log(np)} \le |\Gamma_1^{G\backslash S}(j)| \le 4np\right\}.$$

Note that $G\backslash S \sim \mathcal{G}(n-|S|,p)$ and $(4np)^\ell = o(n)$. Applying Corollary 1 together with union bounds, we get that

$$\mathbb{P}\{\mathcal{A}\} \ge 1 - n|S|\exp\left\{-\Omega\left((np)^2/\log(np)\right)\right\} \ge 1 - n^{\omega(1)}.$$

Define event $\mathcal{B}$ such that for every vertex i, there is at most 1 neighbor j in G such that $|\Gamma_1^{G\backslash S}(j)| \le (np)/\log(np)$. Recall $\mathcal{E}$ is the event that the maximum degree in G is at most $4np$. In view of Lemma 16, we have that $\mathbb{P}\{\mathcal{B}\cap\mathcal{E}|\} \ge 1-n^{-1+o(1)}$.

Recall that $\ell = \lfloor\frac{(1-\epsilon)\log n}{\log(np)}\rfloor$ and $\eta = 4^{2\ell+2}n^{1-2\epsilon}\alpha$. Then for sufficiently large n,

$$\frac{(np)^\ell}{2^{\ell-1}\log(np)} \ge \frac{4\eta}{\alpha}.$$

Hence, on event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{E}$,

$$\sum_{j\in\Gamma_1^G(i)} a_{ij} \ge d_i - 1, \quad \forall i.$$

Let

$$\mathcal{X} = \cup_{i,j}\left(\{a_{ij}=1\}\cap\{b_{ij}=0\}\right)$$

Then on event $\mathcal{X}^c$, for all i,j such that $a_{ij}=1$, $b_{ij}=1$; thus $a_{ij}\le b_{ij}$ for all i,j. Hence, on event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{E}\cap\mathcal{X}^c$, we have

$$\sum_{j\in\Gamma_1^G(i)} b_{ij} \ge d_i - 1, \quad \forall i.$$

It remains to show $\mathbb{P}\{\mathcal{A}\cap\mathcal{B}\cap\mathcal{E}\cap\mathcal{X}^c\}\geq 1-n^{-1+o(1)}$, which further reduces to proving $\mathbb{P}\{\mathcal{X}\}\leq n^{-1+o(1)}$ by the union bound. Note that

$$\begin{aligned}\mathbb{P}\{a_{ij}=1,b_{ij}=0\}&\leq\mathbb{P}\{b_{ij}=0\mid a_{ij}=1\}\\&\leq\mathbb{P}\{\mathrm{Bin}(\lfloor 4\eta/\alpha\rfloor,\alpha)\leq\eta\}\\&\leq e^{-\eta},\end{aligned}$$

where the last inequality follows from the Binomial tail bound (32). By the union bound, we have

$$\mathbb{P}\{\mathcal{X}\}\leq\sum_{i,j}\mathbb{P}\{a_{ij}=1,b_{ij}=0\}\leq n^2e^{-\eta}\leq n^{-2},$$

where the last inequality holds due to the assumption that $\eta\geq 4\log n$.

□

The second lemma is useful to conclude that the test statistic $Z_{u,v}$ given in (17) is small for all distinct vertices u,v.

Lemma 9. *Assume the same setup as Lemma 8. With probability at least $1-4/n$, for all distinct u,v, there exists a constant C depending only on ϵ such that*

$$\sum_{i\in\Gamma_1^G(u)}\sum_{j\in\Gamma_1^G(v)}\mathbf{1}_{\left\{|N_\ell^{G\setminus\{u,v\}}(i)\cap N_\ell^{G\setminus\{u,v\}}(j)\cap I_0|\geq\eta\right\}}\leq C.$$

Proof. For two vertices i,j, define

$$c_{ij}=\mathbf{1}_{\left\{|N_\ell^G(i)\cap N_\ell^G(j)|\geq\eta/(4\alpha)\right\}}$$

and an event

$$\mathcal{C}=\left\{\max_i\sum_j c_{ij}\leq 2n^{4\epsilon}\right\}\cap\left\{\max_j\sum_i c_{ij}\leq 2n^{4\epsilon}\right\}$$

In view of Lemma 14, $\mathbb{P}\{\mathcal{C}\}\geq 1-2/n$.

Define

$$a_{ij}=\mathbf{1}_{\left\{|N_\ell^G(i)\cap N_\ell^G(j)\cap I_0|\geq\eta\right\}}.$$

and an event

$$\mathcal{A}=\left\{\max_i\sum_j a_{ij}\leq 2n^{4\epsilon}\right\}\cap\left\{\max_j\sum_i a_{ij}\leq 2n^{4\epsilon}\right\}$$

Moreover, let

$$\mathcal{Y}=\cup_{i,j}\left[\{c_{ij}=0\}\cap\{a_{ij}=1\}\right]$$

Then on $\mathcal{Y}^c$, for all i,j such that $c_{ij}=0$, $a_{ij}=0$; thus $a_{ij}\leq c_{ij}$ for all i,j. Hence, $\mathcal{C}\cap\mathcal{Y}^c\subset\mathcal{A}$ and thus

$$\mathbb{P}\{\mathcal{A}\}\geq\mathbb{P}\{\mathcal{C}\cap\mathcal{Y}^c\}\geq\mathbb{P}\{\mathcal{C}\}-\mathbb{P}\{\mathcal{Y}\}.$$

Note that

$$\begin{aligned}\mathbb{P}\{c_{ij}=0,a_{ij}=1\}&\leq\mathbb{P}\{a_{ij}=1\mid c_{ij}=0\}\\&\leq\mathbb{P}\{\mathrm{Bin}(\lfloor\eta/4\alpha\rfloor,\alpha)\geq\eta\}\\&\leq e^{-2\eta}\end{aligned}$$

By the union bound, we have

$$\mathbb{P}\{\mathcal{Y}\} \le \sum_{i,j} \mathbb{P}\{c_{ij} = 0, a_{ij} = 1\} \le n^2 e^{-2\eta} \le n^{-6},$$

where the last inequality follows from the assumption that $\eta \ge 4 \log n$. Thus $\mathbb{P}\{\mathcal{A}\} \ge 1 - 3/n$.

Fix a pair of vertices $u \ne v$ in the sequel, and let

$$b_{ij} = \mathbf{1}_{\left\{|N_\ell^{G\backslash\{u,v\}}(i) \cap N_\ell^{G\backslash\{u,v\}}(j) \cap I_0| \ge \eta\right\}}$$

and

$$\mathcal{B}_{u,v} = \left\{\max_i \sum_j b_{ij} \le 2n^{4\epsilon}\right\} \cap \left\{\max_j \sum_i b_{ij} \le 2n^{4\epsilon}\right\}$$

Then by construction, $b_{ij} \le a_{ij}$ and thus $\mathcal{A} \subset \mathcal{B}_{u,v}$.

Let $X_i = G(u,i)$ for $i \in [n]$ and $X_{n+j} = G(v,j)$ for $j \in [n]$. Define

$$R_{u,v} = \sum_{i,j \in [n]\backslash\{u,v\}} b_{ij} X_i X_{n+j},$$

which is a degree-2 polynomial of X_i's. Note that $\{b_{ij}; i,j \in [n]\backslash\{u,v\}\}$ only depends on $G\backslash\{u,v\}$ and hence is independent from X_i's. Moreover, X_i's are i.i.d. Bern(p).

We condition on $\{b_{ij}\}$ such that event $\mathcal{B}_{u,v}$ holds. Let

$$\mu_0 = \mathbb{E}[R_{u,v} \mid b] = p^2 \sum_{ij} b_{ij} \le 2p^2 n^{1+4\epsilon} \le 2n^{-1+6\epsilon}.$$

and

$$\mu_1 = \max\left\{\max_i \mathbb{E}\left[\sum_j b_{ij} X_j \mid b\right], \max_j \mathbb{E}\left[\sum_i b_{ij} X_i \mid b\right]\right\} \le 2pn^{4\epsilon} \le 2n^{-1+5\epsilon}.$$

By a concentration inequality for multivariate polynomials [Vu02, Corollary 4.9], there exists a constant $C > 0$ depending only on ϵ such that

$$\mathbb{P}\{R_{u,v} \ge C \mid b\} \le n^{-3}.$$

Thus $\mathbb{P}\{R_{u,v} \ge C \mid \mathcal{B}_{u,v}\} \le n^{-3}$. Define event $\mathcal{R}_{u,v} = \{R_{u,v} \le C\}$ and $\mathcal{R} = \cap_{u \ne v} \mathcal{R}_{u,v}$. It follows that

$$\mathbb{P}\{\mathcal{R}_{u,v} \cap \mathcal{B}_{u,v}\} \le \mathbb{P}\{\mathcal{R}_{u,v} \mid \mathcal{B}_{u,v}\} \le n^{-3}.$$

Since $\mathcal{A} \subset \mathcal{B}_{u,v}$, it further follows that $\mathbb{P}\{\mathcal{R}_{u,v} \cap \mathcal{A}\} \le n^{-3}$. By a union bound over u, v, we have $\mathbb{P}\{\mathcal{R} \cap \mathcal{A}\} \le n^{-1}$. Hence, $\mathbb{P}\{\mathcal{R}\} \le \mathbb{P}\{\mathcal{R} \cap \mathcal{A}\} + \mathbb{P}\{\mathcal{A}^c\} \le 4/n$.

□

With Lemma 8 and Lemma 9, we are ready to finish the proof of Theorem 3.

Proof of Theorem 3. Recall that τ is given in (14) and the definition of high-degree vertices. We first prove that Algorithm 2 correctly matches the high-degree vertices in $G_1^* \wedge G_2$ with high probability.

Recall the definition of Z give in (17). Applying Lemma 8 with $G = G_1^* \wedge G_2$, we get that with high probability, for all high-degree vertices u,

$$Z_{u,u} \ge \tau - 1 = \frac{nps^2}{\log(nps^2)} - 1.$$

Moreover, by definition,

$$\begin{aligned} w_{i,j}^{u,v} &\le \left| \left\{ k \in I_0 : \pi_0(k) \in N_\ell^{G_1 \setminus \{u,v\}}(i),\ k \in N_\ell^{G_2 \setminus \{u,v\}}(j) \right\} \right| \\ &\le \left| N_\ell^{G_1^* \vee G_2 \setminus \{u,v\}}(i) \cap N_\ell^{G_1^* \vee G_2 \setminus \{u,v\}}(j) \cap I_0 \right|. \end{aligned}$$

Applying Lemma 9 with $G = G_1^* \vee G_2$, we get that with high probability,

$$Z_{u,v} \le C, \quad \forall u \neq v$$

for a constant $C > 0$ only depending on ϵ. Since for sufficiently large n, $\tau \ge C + 1$, it follows that Algorithm 2 correctly matches all high-degree vertices with high probability.

The proof of correctness for matching low-degree vertices is the same as Algorithm 1 and thus omitted.

□

Appendix A Proof of Theorem 2

Proof of Theorem 2. Suppose $nps^2 - \log n = c$ for $c < +\infty$. Since $G_1^* \wedge G_2 \sim \mathcal{G}(n, ps^2)$, classical random graph theory shows that the distribution of the number of isolated vertices in $G_1^* \wedge G_2$ converges to $\mathrm{Pois}(e^{-c})$, see, e.g., [Bol01, Theorem 3.1]. Let $\mathcal{F}_1$ denote the event that there are at least two isolated vertices in $G_1^* \wedge G_2$. Then $\mathbb{P}\{\mathcal{F}_1\} = \Omega(1)$.

Let $\mathcal{F}_2$ denote the event that there are at least two isolated vertices that are unseeded in $G_1^* \wedge G_2$. Since each vertex is seeded with probability α independently across different vertices and from the graphs G_1 and G_2, it follows that $\mathbb{P}\{F_2\} \ge \mathbb{P}\{F_1\}(1-\alpha)^2 = \Omega\left((1-\alpha)^2\right)$.

Since the prior distribution of π^* is uniform, the maximum likelihood estimator $\widehat{\pi}_{\mathrm{ML}}$ minimizes the error probability $\mathbb{P}\{\widehat{\pi} \neq \pi^*\}$ among all possible estimators and thus we only need to find when MLE fails.

Recall that I_0 is the seed set. Let $\mathcal{S}$ denote the set of all possible permutations π such that $\pi(i) = \pi^*(i)$ for $i \in I_0$. Under the seeded model $\mathcal{G}(n, p; s, \alpha)$, the maximum likelihood estimator $\widehat{\pi}_{\mathrm{ML}}$ is given by the minimizer of the (restricted) quadratic assignment problem, namely,

$$\widehat{\pi}_{\mathrm{ML}} \in \arg\min_{\pi \in \mathcal{S}} \min \|G_1 - \Pi G_2 \Pi^\top\|_F^2,$$

where Π is the permutation matrix corresponding to permutation π; or equivalently,

$$\widehat{\pi}_{\mathrm{ML}} \in \arg\max_{\pi \in \mathcal{S}} \left\langle G_1, \Pi G_2 \Pi^\top \right\rangle.$$

Let I denote the union of the initial seed set and the set of all non-isolated vertices in $G_1^* \wedge G_2$. Then I^c is the set of isolated vertices that are unseeded in $G_1^* \wedge G_2$. Let $\widetilde{\mathcal{S}}$ denote the set of all possible permutations π such that $\pi(i) = \pi^*(i)$ for $i \in I$. Then $\pi^* \in \widetilde{\mathcal{S}} \subset \mathcal{S}$. Note that for any $\pi \in \widetilde{\mathcal{S}}$, we have

$$\begin{aligned} \left\langle G_1, \Pi G_2 \Pi^\top \right\rangle &\ge \sum_{(i,j) \in I \times I} G_1(\pi(i), \pi(j)) G_2(i,j) \\ &\overset{(a)}{=} \sum_{(i,j) \in I \times I} G_1(\pi^*(i), \pi^*(j)) G_2(i,j) \\ &= \sum_{(i,j)} G_1(\pi^*(i), \pi^*(j)) G_2(i,j), \end{aligned}$$

where (a) follows from $\pi(i) = \pi^*(i)$ for $i \in I$; the last equality holds due to $G_1(\pi^*(i), \pi^*(j))G_2(i,j) = 0$ for all $(i,j) \notin I \times I$. Hence, there at at least $|I^c|! - 1$ different permutations in $\widehat{\mathcal{S}}$ whose likelihood is at least as large as the ground truth π^*, and hence the MLE is correct with probability at most $1/(|I^c|! - 1)$. Note that on event $\mathcal{F}_2$, $|I^c| \geq 2$; hence, MLE is correct with probability at most $1/2$. In conclusion, MLE is correct with probability at most $(1/2)\mathbb{P}\{\mathcal{F}_2\} = \Omega((1-\alpha)^2)$.

□

Appendix B Proof of Lemma 1

Proof. Claim (i): For each vertex i, its degree $d_i \sim \text{Binom}(n-1, p)$. By the union bound, the probability that G has an isolated vertex is

$$n(1-p)^{n-1} \leq ne^{-(n-1)p} = o(1),$$

where the last equality holds due to the assumption that $np - \log n \to +\infty$.

Claim (ii): Fix any pair of two distinct vertices i, j, define

$$\mathcal{E}_{ij} = \{G(i,j) = 1\} \cap \{d_i \leq \tau\} \cap \{d_j \leq \tau\}.$$

It suffices to show

$$\mathbb{P}\{\cup_{i \neq j} \mathcal{E}_{ij}\} \leq n^{-1+o(1)}.$$

Note that

$$\begin{aligned}\mathbb{P}\{d(i) \leq \tau, d(j) \leq \tau | G(i,j) = 1\} &= (\mathbb{P}\{\text{Bin}(n-2,p) \leq \tau - 1\})^2 \\ &\leq (\mathbb{P}\{\text{Bin}(n-2,p) \leq \tau\})^2\end{aligned}$$

In view of Binomial tail bounds given in Theorem 6 and $\tau = o(np)$, we have that

$$\mathbb{P}\{\text{Bin}(n-2,p) \leq \tau\} \leq \exp\left(-(n-2)p\left(1 - \sqrt{\frac{\tau}{(n-2)p}}\right)^2\right) = \exp\left(-(1-o(1))np\right).$$

Combining the last two displayed equations yields that

$$\begin{aligned}\mathbb{P}\{\mathcal{E}_{ij}\} &= \mathbb{P}\{G(i,j) = 1\}\,\mathbb{P}\{d(i) \leq \tau, d(j) \leq \tau | G(i,j) = 1\} \\ &\leq p \exp\left(-2(1-o(1))np\right)\end{aligned}$$

By the union bound,

$$\begin{aligned}\mathbb{P}\{\cup_{i \neq j} E_{ij}\} &\leq n^2 \mathbb{P}\{E_{ij}\} \\ &\leq n^2 p \exp\left(-2(1-o(1))np\right) = n^{-1+o(1)},\end{aligned}$$

where the last equality holds due to $np - \log n \to +\infty$. □

Appendix C Proof of Lemma 2

Proof. Let d_i denote the degree of vertex i in $G_1^* \wedge G_2$ and A denote the adjacency matrix of $G_1^* \vee G_2$. For every pair of three distinct vertices i, j, k, define

$$\mathcal{F}_{ijk} = \{A_{ik} = 1, A_{jk} = 1\} \cap \{d_i \le \tau\} \cap \{d_j \le \tau\}.$$

It suffices to show that $\mathbb{P}\{\cup_{i,j,k} F_{ij}\} \le n^{-1+o(1)}$. Since $G_1^* \vee G_2 \sim \mathcal{G}(n, ps(2-s))$, it follows that

$$\mathbb{P}\{A_{ik} = 1, A_{jk} = 1\} = \mathbb{P}\{A_{ik} = 1\}\mathbb{P}\{A_{jk} = 1\} = (ps(2-s))^2 \le p^2.$$

Moreover, since $G_1^* \wedge G_2 \sim \mathcal{G}(n, ps^2)$, it follows that

$$\mathbb{P}\{\{d_i \le \tau\} \cap \{d_j \le \tau\} \mid A_{ik} = 1, A_{jk} = 1\} \le \left(\mathbb{P}\left\{\mathrm{Binom}(n-3, ps^2) \le \tau\right\}\right)^2.$$

In view of Binomial tail bound (32) and $\tau = o(nps^2)$, we have that

$$\begin{aligned}\mathbb{P}\left\{\mathrm{Binom}(n-3, ps^2) \le \tau\right\} &\le \exp\left(-(n-3)ps^2\left(1 - \sqrt{\frac{\tau}{(n-3)ps^2}}\right)^2\right) \\ &= \exp\left(-nps^2\left(1 - o(1)\right)\right)\end{aligned}$$

It follows that

$$\mathbb{P}\{\mathcal{F}_{ijk}\} \le p^2 \exp\left(-2nps^2\left(1 - o(1)\right)\right)$$

By the union bound, we have that

$$\mathbb{P}\{\cup_{i,j,k} F_{ij}\} \le n^3 p^2 \exp\left(-2nps^2\left(1 - o(1)\right)\right) = n^{-1+o(1)}.$$

where the last equality holds due to $nps^2 \ge \log n$ and $\log(np) = o(nps^2)$. □

Appendix D Neighborhood Exploration in $\mathcal{G}(n, p)$

Throughout this section, we assume graph $G \sim \mathcal{G}(n, p)$ with $np \ge \log n$. We first claim that the max degree in G is at most $4np$ with probability at least $1 - 1/n$.

Lemma 10. *Assume graph $G \sim \mathcal{G}(n, p)$ with $np \ge \log n$. Let*

$$\mathcal{E} = \left\{\max_{v \in V(G)} d_v \le 4np\right\}. \tag{23}$$

Then

$$\mathbb{P}\{\mathcal{E}\} \ge 1 - n^{-1}.$$

Proof. By the Binomial tail bound (33),

$$\mathbb{P}\{d_i \ge 4np\} = \mathbb{P}\{\mathrm{Binom}(n-1, p) \ge 4np\} \le \exp(-2np).$$

The proof follows by the union bound and the assumption that $np \ge \log n$. □

We fix a vertex u throughout this section, and abbreviate $\Gamma_k^G(u)$ as $\Gamma_k(u)$ and $N_k^G(u)$ as $N_k(u)$ for simplicity. We are interested in studying the growth of $|\Gamma_k(u)|$ as k increases. Note that $|\Gamma_1(u)|$ is the degree d_u of vertex u in G. Since the average degree is $(n-1)p$, we expect typically $|\Gamma_k(u)|$ grows as $(np)^k$. This is indeed true in the dense regime with $np \ge n^\epsilon$.

D.1 Dense Regime

The following lemma is adapted from [Bol01, Lemma 10.9].

Lemma 11. *Suppose $np \geq n^{\epsilon}$ for an arbitrarily small constant $\epsilon > 0$ and d is chosen such that*

$$(np)^{d-1} \leq \frac{n}{8} \quad \text{and} \quad (np)^{d} \geq n \log n$$

If n is sufficiently large, then with probability at least $1 - n^{-10}$, the following claims hold:

(i) For every vertex u,

$$\left|\Gamma_k(u) - (np)^k\right| \leq \frac{1}{4}(np)^k. \quad \forall 0 \leq k \leq d-1.$$

(ii) For every two distinct vertices u and v,

$$|N_{d-1}(u) \cap N_{d-1}(v)| \leq 8n^{2d-3}p^{2d-2}.$$

Lemma 11 also upper bounds $|\Gamma_{d-1}(u) \cap \Gamma_{d-1}(v)|$ for two distinct vertices u, v by $8p^{2d-2}n^{2d-3}$. To see this intuitively, note that in the dense regime, $\Gamma_{d-2}(u) \cap \Gamma_{d-2}(v)$ is typically of a much smaller size comparing to either $\Gamma_{d-2}(u)$ or $\Gamma_{d-2}(v)$. Hence, the majority of vertices w in $\Gamma_{d-1}(u) \cap \Gamma_{d-1}(v)$ are connected to some vertex in $\Gamma_{d-2}(u) \setminus \Gamma_{d-2}(v)$ and to some vertex in $\Gamma_{d-2}(v) \setminus \Gamma_{d-2}(u)$. For a given vertex $w \notin N_{d-2}(u) \cup N_{d-2}(v)$, since $|\Gamma_{d-2}(u) \setminus \Gamma_{d-2}(v)| \leq |\Gamma_{d-2}(u)| \leq 2(np)^{d-2}$ and similarly for $|\Gamma_{d-2}(v) \setminus \Gamma_{d-2}(u)|$, w connects to some vertex in $\Gamma_{d-2}(u) \setminus \Gamma_{d-2}(v)$ with probability at most $2p(np)^{d-2}$, and connects to some vertex in $\Gamma_{d-2}(v) \setminus \Gamma_{d-2}(u)$ independently with probability $2p(np)^{d-2}$. Moreover, there are at most n such potential vertices w to consider. Hence, we expect $|\Gamma_{d-1}(u) \cap \Gamma_{d-1}(v)|$ to be smaller than $2n[2p(np)^{d-2}]^2 = 8p^{2d-2}n^{2d-3}$.

D.2 Sparse Regime

In contrast, in the sparse regime where

$$np - \log n \to +\infty.$$

there exist vertices with small degrees, i.e., $|\Gamma_1(u)|$ is much smaller than np. Hence, we cannot expect $|\Gamma_k(u)|$ grows like $(np)^k$ for all vertices u. Nevertheless, the following lemma shows that conditional on $|\Gamma_1(u)|$ is large, then $|\Gamma_k(u)| \asymp (np)|\Gamma_{k-1}(u)|$ for all $2 \leq k \leq d$ for some d with high probability.

Lemma 12. *Suppose*

$$np \geq \log n \quad \text{and} \quad p(4np)^{d-1} = o(1). \tag{24}$$

Let u be a fixed vertex. For each $1 \leq k \leq d$, define

$$\mathcal{Q}_k = \left\{ |\Gamma_k(u)| \in I_k = \left[\tau \left(\frac{np}{2}\right)^{k-1}, (4np)^k \right] \right\}$$

for $1 \leq \tau \leq np$. Then for $2 \leq k \leq d$,

$$\mathbb{P}\left\{ \mathcal{Q}_k \mid \mathcal{Q}_1, \ldots, \mathcal{Q}_{k-1} \right\} \geq 1 - \exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1} \right) \right).$$

It readily follows that

$$\mathbb{P}\left\{ \mathcal{Q}_d \cap \mathcal{Q}_{d-1} \cap \cdots \cap \mathcal{Q}_2 \mid \mathcal{Q}_1 \right\} \geq 1 - \exp\left(-\Omega\left(\tau np\right)\right).$$

Proof. Fix $2 \le k \le d$. Conditional on $\Gamma_{k-1}(u)$ and $N_{k-1}(u)$, the probability of a given vertex $v \notin N_{k-1}(u)$ being connected to some vertices in $\Gamma_{k-1}(u)$ is

$$p_k \triangleq 1-(1-p)^{|\Gamma_{k-1}(u)|}.$$

Therefore, conditional on $|\Gamma_{k-1}(u)|$ and $|N_{k-1}(u)|$,

$$|\Gamma_k(u)| \sim \mathrm{Bin}\left(n-|N_{k-1}(u)|, p_k\right)$$

Note that conditional on $\mathcal{Q}_1, \ldots, \mathcal{Q}_{k-1}$,

$$|N_{k-1}(u)| = \sum_{i=0}^{k-1} |\Gamma_i(u)| \le \sum_{i=0}^{k-1} (4np)^i = \frac{(4np)^k - 1}{4np-1} = o(n),$$

where the last equality holds due to the assumption (24) and $k \le d$. Moreover, in view of the assumption (24), conditional on $\mathcal{Q}_1, \ldots, \mathcal{Q}_{k-1}$,

$$(1-o(1))\, p\tau \left(\frac{np}{2}\right)^{k-2} \le p_k \le p(4np)^{k-1}.$$

Hence, for $2 \le k \le d$,

$$\begin{aligned}
\mathbb{P}\left\{|\Gamma_k(u)| \notin I_k \mid \mathcal{Q}_1, \ldots, \mathcal{Q}_k\right\} &\le \mathbb{P}\left\{\mathrm{Bin}\left(n-o(n), (1-o(1))\, p\tau\left(\frac{np}{2}\right)^{k-2}\right) \le \tau\left(\frac{np}{2}\right)^{k-1}\right\} \\
&\quad + \mathbb{P}\left\{\mathrm{Bin}\left(n, p(4np)^{k-1}\right) \ge (4np)^k\right\} \\
&\le \exp\left(-\Omega\left(\tau\left(\frac{np}{2}\right)^{k-1}\right)\right) + \exp\left(-4^{k-1}(np)^k\right) \\
&\le \exp\left(-\Omega\left(\tau\left(\frac{np}{2}\right)^{k-1}\right)\right).
\end{aligned}$$

Finally, we note that

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{Q}_d \cap \mathcal{Q}_{d-1} \cap \cdots \cap \mathcal{Q}_2 \mid \mathcal{Q}_1\right\} &= \mathbb{P}\left\{\mathcal{Q}_2 \mid \mathcal{Q}_1\right\} \mathbb{P}\left\{\mathcal{Q}_3 \mid \mathcal{Q}_1, \mathcal{Q}_2\right\} \cdots \mathbb{P}\left\{\mathcal{Q}_d \mid \mathcal{Q}_1, \ldots, \mathcal{Q}_{d-1}\right\} \\
&\ge \prod_{i=0}^{d-1} \left(1-\exp\left(-\Omega\left(\tau\left(\frac{np}{2}\right)^{k-1}\right)\right)\right) \\
&\ge 1-\sum_{i=0}^{d-1} \exp\left(-\Omega\left(\tau\left(\frac{np}{2}\right)^{k-1}\right)\right) \\
&\ge 1-\exp\left(-\Omega\left(\tau np\right)\right).
\end{aligned}$$

□

With Lemma 12, we have the following immediate corollary.

Corollary 1. *Suppose (24) holds. Define event*

$$\mathcal{Q} = \left\{|\Gamma_k(u)| \in I_k,\ \forall 1 \le k \le d,\ \forall u \text{ s.t. } \tau \le |\Gamma_1(u)| \le 4np\right\}$$

Then

$$\mathbb{P}\left\{\mathcal{Q}\right\} \ge 1-n\exp\left(-\Omega\left(\tau np\right)\right).$$

Proof. Note that

$$\mathcal{Q}^c = \cup_u \left(\{\tau \le |\Gamma_1(u)| \le 4np\} \cap \{|\Gamma_k(u)| \notin I_k,\ \forall 1 \le k \le d\} \right).$$

Hence, it follows from the union bound that

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{Q}^c\right\} &\le \sum_u \mathbb{P}\left\{ \{\tau \le |\Gamma_1(u)| \le 4np\} \cap \{|\Gamma_k(u)| \notin I_k,\ \forall 1 \le k \le d\} \right\} \\
&\le \sum_u \mathbb{P}\left\{ |\Gamma_k(u)| \notin I_k,\ \forall 1 \le k \le d \mid \tau \le |\Gamma_1(u)| \le 4np \right\} \\
&\le n \exp\left(-\Omega\left(\tau np\right)\right),
\end{aligned}$$

where the last inequality follows from Lemma 12. □

Next, we upper bounds $|N_d(u) \cap N_d(v)|$ for two distinct vertices u, v in the sparse regime. We need to introduce

$$\Gamma^*_{k,\ell}(u,v) = \{w \in \Gamma_k(u) \cap \Gamma_\ell(v) : \Gamma_1(w) \cap (\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)) \neq \emptyset,\ \Gamma_1(w) \cap (\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)) \neq \emptyset\}$$

and we abbreviate $\Gamma^*_{k,k}(u,v)$ as $\Gamma^*_k(u,v)$ for simplicity. By definition, for any $d \ge 1$,

$$\Gamma_d(u) \cap \Gamma_d(v) \subset \cup_{k=1}^d \Gamma_{d-k}\left(\Gamma^*_k(u,v)\right).$$

and

$$N_d(u) \cap N_d(v) \subset \cup_{\ell=-d}^d \cup_{k=0}^d N_{d-k-\max\{\ell,0\}}\left(\Gamma^*_{k+\ell,k}(u,v)\right).$$

The following lemma gives an upper bound to $\left|\Gamma^*_{k,\ell}(u,v)\right|$ in high probability.

Lemma 13. *For two distinct vertices u, v, define*

$$\Delta_{k,\ell} = \left\{ |\Gamma_{k-1}(u)| \le (4np)^{k-1},\ |\Gamma_{\ell-1}(v)| \le (4np)^{\ell-1} \right\}.$$

For all $k \ge 1$,

$$\mathbb{P}\left\{ \left|\Gamma^*_{k,\ell}(u,v)\right| \ge \gamma_{k+\ell} \mid \Delta_{k,\ell} \right\} \le \frac{1}{n^8}, \tag{25}$$

where

$$\gamma_k = \begin{cases} 24 \log n & \text{if } np^2(4np)^{k-2} \le 4 \log n \\ 4np^2(4np)^{k-2} & \text{o.w.} \end{cases} \tag{26}$$

Proof. Conditional on $\mathcal{N}_{k-1}(u), \Gamma_{k-1}(u)$ and $\mathcal{N}_{\ell-1}(v), \Gamma_{\ell-1}(v)$, the probability that a vertex $w \notin \mathcal{N}_{k-1}(u) \cup \mathcal{N}_{\ell-1}(v)$ being connected to some vertex in $\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)$ is

$$1 - (1-p)^{|\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)|} \le p|\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)| \le p|\Gamma_{k-1}(u)|.$$

Similarly, the probability that w is connected to some vertex in $\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)$ is

$$1 - (1-p)^{|\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)|} \le p|\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)| \le p|\Gamma_{\ell-1}(v)|.$$

Since $\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)$ is disjoint from $\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)$, the probability that $w \in \Gamma^*_{u,v}$ is at most $p^2|\Gamma_{k-1}(u)||\Gamma_{\ell-1}(v)|$. Moreover, there are at most n vertices $w \notin \mathcal{N}_{k-1}(u) \cup \mathcal{N}_{\ell-1}(v)$. Hence,

$$\mathbb{P}\left\{\left|\Gamma^*_{k,\ell}(u,v)\right| \geq \gamma_{k+\ell} \mid \Delta_{k,\ell}\right\} \leq \mathbb{P}\left\{\mathrm{Bin}\left(n, p^2(4np)^{k+\ell-2}\right) \geq \gamma_{k+\ell}\right\}.$$

If $np^2(4np)^{k+\ell-2} \leq 4\log n$, then by the choice of $\gamma_{k+\ell} = 24\log n$, we have $\gamma_{k+\ell} \geq 6np^2(4np)^{k+\ell-2}$. It follows from (34) that

$$\mathbb{P}\left\{\mathrm{Bin}\left(n, p^2(4np)^{k+\ell-2}\right) \geq \gamma_{k+\ell}\right\} \leq 2^{-\gamma_{k+\ell}} = 2^{-24\log n} \leq \frac{1}{n^8}.$$

If $np^2(4np)^{k+\ell-2} \geq 4\log n$, then by the choice of $\gamma_{k+\ell} = 4np^2(4np)^{k+\ell-2}$, it follows from (33) that

$$\mathbb{P}\left\{\mathrm{Bin}\left(n, p^2(4np)^{k+\ell-2}\right) \geq \gamma_{k+\ell}\right\} \leq \exp\left(-2np^2(4np)^{k+\ell-2}\right) \leq \frac{1}{n^8}.$$

□

With Lemma 13, we are ready to upper bound $|N_d(u) \cap N_d(v)|$ for d large enough.

Lemma 14. *For a given small constant $\epsilon > 0$, choose an integer $1 \leq d \leq n$ such that*

$$(4np)^d \geq n^{1-\epsilon}$$

For each vertex u, define event

$$\mathcal{R}_u = \left\{\sum_v \mathbf{1}_{\left\{|N_d(u)\cap N_d(v)| > 4^{2d+1}p^{2d}n^{2d-1}\right\}} \leq 2n^{4\epsilon}\right\}$$

and $\mathcal{R} = \cap_u \mathcal{R}_u$. Then

$$\mathbb{P}\{\mathcal{R}\} \geq 1 - 2n^{-1}. \tag{27}$$

Proof. Define an event

$$\mathcal{A} = \cap_{u\neq v} \cap_{1\leq k\leq d} \cap_{1\leq k\leq \ell} \left\{\left|\Gamma^*_{k,\ell}(u,v)\right| \leq \gamma_{k+\ell}\right\}$$

Recall $\mathcal{E}$ defined in (23). Note that

$$(\mathcal{A}\cap\mathcal{E})^c = (\mathcal{A}^c \cap \mathcal{E}) \cup \mathcal{E}^c.$$

Therefore,

$$\mathbb{P}\left\{(\mathcal{A}\cap\mathcal{E})^c\right\} \leq \mathbb{P}\left\{\mathcal{A}^c \cap \mathcal{E}\right\} + \mathbb{P}\left\{\mathcal{E}^c\right\}.$$

Note that $\mathbb{P}\{\mathcal{E}^c\} \leq 1/n$. Moreover,

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{A}^c \cap \mathcal{E}\right\} &\leq \sum_{u\neq v}\sum_{1\leq k\leq d}\sum_{1\leq \ell\leq d} \mathbb{P}\left\{\left\{\left|\Gamma^*_{k,\ell}(u,v)\right| \geq \gamma_{k+\ell}\right\} \cap \mathcal{E}\right\} \\
&\overset{(a)}{\leq} \sum_{u\neq v}\sum_{1\leq k\leq d}\sum_{1\leq \ell\leq d} \mathbb{P}\left\{\left\{\left|\Gamma^*_{k,\ell}(u,v)\right| \geq \gamma_{k+\ell}\right\} \cap \Delta_{k,\ell}\right\} \\
&\leq \sum_{u\neq v}\sum_{1\leq k\leq d}\sum_{1\leq \ell\leq d} \mathbb{P}\left\{\left|\Gamma^*_{k,\ell}(u,v)\right| \geq \gamma_{k+\ell} \mid \Delta_{k,\ell}\right\} \leq n^{-4},
\end{aligned}$$

where (a) follows from $\mathcal{E} \subset \Delta_k$ and the last inequality holds in view of Lemma 13 and $d \le n$. Therefore, $\mathbb{P}\{(\mathcal{A} \cap \mathcal{E})^c\} \le 2/n$.

To prove the lemma, it suffices to argue that $\mathcal{A} \cap \mathcal{E} \subset \mathcal{R}$. To see this, let us assume that $\mathcal{A} \cap \mathcal{E}$ holds in the sequel. Note that

$$N_d(u) \cap N_d(v) \subset \cup_{\ell=-d}^{d} \cup_{k=0}^{d} N_{d-k-\max\{\ell,0\}} \left(\Gamma^*_{k+\ell,k}(u,v)\right).$$

It follows that

$$|N_d(u) \cap N_d(v)| \le \sum_{\ell=-d}^{d} \sum_{k=0}^{d} \left|\Gamma^*_{k+\ell,k}(u,v)\right| (4np)^{d-k-\max\{\ell,0\}}$$

Set k_0

$$k_0 = \left\lfloor \frac{2\epsilon \log n}{\log(4np)} \right\rfloor$$

Then

$$|N_{2k_0}(u)| \le \sum_{k=0}^{2k_0} (4np)^k = \frac{(4np)^{2k_0+1} - 1}{4np - 1} \le 2(4np)^{2k_0} \le 2n^{4\epsilon},$$

where the second-to-the-last inequality holds due to $2np \ge 1$. Note that for all $v \notin N_{2k_0}(u)$, we have

$$|\Gamma^*_{k,\ell}(u,v)| = 0, \quad \forall 0 \le k + \ell \le 2k_0$$

and thus

$$\begin{aligned}
|N_d(u) \cap N_d(v)| &\le \sum_{\ell=-d}^{d} \sum_{k=0}^{d} \mathbf{1}_{\{0 \le k+\ell \le d\}} \mathbf{1}_{\{2k+\ell \ge 2k_0+1\}} \gamma_{2k+\ell} (4np)^{d-k-\max\{\ell,0\}} \\
&\le \sum_{\ell=-d}^{d} \sum_{k=0}^{d} \mathbf{1}_{\{0 \le k+\ell \le d\}} \mathbf{1}_{\{2k+\ell \ge 2k_0+1\}} \left(24 \log n + 4np^2 (4np)^{2k+\ell-2}\right) (4np)^{d-k-\max\{\ell,0\}} \\
&\le 192 \log n (4np)^{d-k_0-1/2} + 32np^2 (4np)^{2d-2} \\
&\le 64np^2 (4np)^{2d-2} = 4^{2d+1} p^{2d} n^{2d-1},
\end{aligned}$$

where the last inequality holds due to $(4np)^{d+k_0+1/2} \ge 6n \log n$ for n sufficiently large. Hence, for every u,

$$\sum_{v} \mathbf{1}_{\left\{|N_d(u) \cap N_d(v)| > (4)^{2d+1} p^{2d} n^{2d-1}\right\}} \le |N_{2k_0}(u)| \le 2n^{4\epsilon}.$$

As a consequence, $\mathcal{A} \cap \mathcal{E} \subset \mathcal{R}$. □

D.3 Graph Branching in Sparse Regime

In this subsection, we describe a branching process to explore the vertices in $N_k(u)$. See, e.g., [AS08, Section 11.5] for a reference.

Definition 3 (Graph Branching Process)**.** *We begin with u and apply breadth-first-search to explore the vertices in $N_k(u)$. In this process, all vertices will be "live", "dead", or "neutral". The live vertices will be contained in a queue. Initially, at time 0, u is live and the queue consists of only u, and all the other vertices are neutral. At each time step t, a live vertex v is popped from the head of the queue, and we check all pairs $\{v, w\}$ for all neutral vertices w for adjacency. The poped vertex v is now dead and those neutral vertices w adjacent to v are added to the end of the queue (in an arbitrary order) and now are live. The process ends when the queue is empty.*

Note that such a branching process constructs a tree $T(u)$ rooted at u. In particular, at each time step, those neutral vertices w adjacent to the poped vertex v can be viewed as children of v. For each vertex v in $T(u)$, abusing notation slightly, we let $T_k(v)$ denote the subtree rooted at v of depth k in $T(u)$ and $\Pi_k(v)$ denote the set of vertices at distance k from root v in subtree $T_k(v)$. Note that by construction, $\Pi_k(u) = \Gamma_k(u)$ for root u.

We are interested in bounding $|\Pi_k(v)|$ for each children v of root u. The following lemma shows that with high probability, for all childen v of root u such that $|\Pi_1(v)| \geq \tau$, $|\Pi_k(v)|$ grows at least as $\tau\,(np/2)^{k-1}$.

Lemma 15. *Let u be the root vertex and $1 \leq \tau \leq np$. Define*

$$\mathcal{F}_1 = \{|\Pi_1(u)| \leq 4np\} \cap \{|\Pi_1(v)| \leq 4np, \forall v \in \Pi_1(u)\},$$

and for each $2 \leq k \leq d$ define

$$\mathcal{F}_k = \left\{|\Pi_k(v)| \leq (4np)^k, \forall v \in \Pi_1(u)\right\} \cap \left\{|\Pi_k(v)| \geq \tau\,(np/2)^{k-1}, \forall v \in \Pi_1(u) \text{ s.t. } |\Pi_1(v)| \geq \tau\right\}$$

Suppose

$$np \geq \log n \quad \text{and} \quad (4np)^{d+1} = o(n). \tag{28}$$

Then for $2 \leq k \leq d$,

$$\mathbb{P}\{\mathcal{F}_k \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}\} \geq 1 - 8np \exp\left(-\Omega\left(\tau\left(\frac{np}{2}\right)^{k-1}\right)\right).$$

It readily follows that

$$\mathbb{P}\{\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2 \mid \mathcal{F}_1\} \geq 1 - 8np \exp\left(-\Omega\left(\tau np\right)\right).$$

Moreover, by letting

$$\mathcal{A}_u = (\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \mathcal{F}_2) \cup \mathcal{F}_1^c,$$

we have

$$\mathbb{P}\{\mathcal{A}_u^c\} \leq 8np \exp\left(-\Omega\left(\tau np\right)\right).$$

Proof. Fix $2 \leq k \leq d$. Suppose the neighbors of root vertex u are added to the queue in the order of $v_1, v_2, \ldots, v_{d_u}$, where $d_u = |\Pi_1(u)|$. Then by the branching process aforementioned, $\Pi_k(v_1), \ldots, \Pi_k(v_{i-1})$ are revealed before $\Pi_k(v_i)$.

Fix $1 \leq i \leq d_u$ and define

$$\mathcal{F}_{k,i} = \left\{|\Pi_k(v_j)| \leq (3np)^k,\ \forall j \in [i]\right\} \cap \left\{|\Pi_k(v_j)| \geq \tau\left(\frac{np}{2}\right)^{k-1}, \forall j \in [i] \text{ s.t. } |\Pi_1(v_j)| \geq \tau\right\}.$$

Then $\mathcal{F}_k = \mathcal{F}_{k,d_u}$.

Conditional on $\Pi_{k-1}(v_i)$, the probability of a given neutral vertex w being connected to some vertices in $\Pi_{k-1}(v_i)$ is

$$p_k \triangleq 1 - (1-p)^{|\Pi_{k-1}(v_i)|} \leq p|\Pi_{k-1}(v_i)|.$$

On the one hand, there are at most n neutral vertices. Therefore, conditional on $|\Pi_{k-1}(v_i)|$, $|\Pi_k(v_i)|$ is stochastically dominated by $\mathrm{Bin}\,(n,\ p\,|\Pi_{k-1}(v_i)|)$ and hence

$$\begin{aligned}\mathbb{P}\left\{|\Pi_k(v_i)| \geq (4np)^k \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}, \mathcal{F}_{k,i-1}\right\} &\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, p(4np)^{k-1}\right) \geq (4np)^k\right\} \\ &\leq \exp\left(-4^{k-1}(np)^k\right), \end{aligned} \tag{29}$$

where the last inequality follows from the Binomial tail bound (33).

On the other hand, in view of assumption (28), conditional on $\mathcal{F}_1, \ldots, \mathcal{F}_{k-1}, \mathcal{F}_{k,i-1}$ there are at least

$$n - 1 - \sum_{i=1}^{d_u} \sum_{\ell=0}^{k-1} |\Pi_\ell(v_i)| - \sum_{j=1}^{i-1} |\Pi_k(v_j)| \geq n - 1 - 4np \sum_{\ell=0}^{k} (4np)^\ell = n - \frac{(4np)^{k+2} - 1}{4np - 1} = n - o(n)$$

neutral vertices to be connected to some vertices in $\Pi_{k-1}(v_i)$, and for each v_i such that $|\Pi_1(v_i)| \geq \tau$,

$$p_k = 1 - (1-p)^{|\Pi_{k-1}(v_i)|} \geq (1 - o(1))\, p\tau \left(\frac{np}{2}\right)^{k-2}.$$

Therefore, conditional on $\mathcal{F}_1, \ldots, \mathcal{F}_{k-1}, \mathcal{F}_{k,i-1}$, $|\Pi_k(v_i)|$ is stochastically lower bounded by

$$\mathrm{Bin}\left(n - o(n), (1 - o(1))\, p\tau \left(\frac{np}{2}\right)^{k-2}\right)$$

and hence for $2 \leq k \leq d$,

$$\begin{aligned}
&\mathbb{P}\left\{|\Pi_k(v_i)| \geq \tau \left(\frac{np}{2}\right)^{k-1} \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}, \mathcal{F}_{k,i-1}\right\} \\
&\leq \mathbb{P}\left\{\mathrm{Bin}\left(n - o(n), (1 - o(1))\, p\tau \left(\frac{np}{2}\right)^{k-2}\right) \leq \tau \left(\frac{np}{2}\right)^{k-1}\right\} \\
&\leq \exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right). \qquad (30)
\end{aligned}$$

Combining (29) and (30) yields that

$$\mathbb{P}\left\{\mathcal{F}_{k,i} \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}\right\} \geq \mathbb{P}\left\{\mathcal{F}_{k,i-1} \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}\right\} \left(1 - 2\exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right)\right).$$

Therefore,

$$\mathbb{P}\left\{\mathcal{F}_k \mid \mathcal{F}_1, \ldots, \mathcal{F}_{k-1}\right\} \geq 1 - 8np\exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right).$$

Finally, we note that

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2 \mid \mathcal{F}_1\right\} \\
&= \mathbb{P}\left\{\mathcal{F}_2 \mid \mathcal{F}_1\right\} \mathbb{P}\left\{\mathcal{F}_3 \mid \mathcal{F}_1, \mathcal{F}_2\right\} \cdots \mathbb{P}\left\{\mathcal{F}_d \mid \mathcal{F}_1, \ldots, \mathcal{F}_{d-1}\right\} \\
&\geq \prod_{k=2}^{d} \left(1 - 8np\exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right)\right) \\
&\geq 1 - 8np \sum_{k=2}^{d} \exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right) \\
&\geq 1 - 8np\exp\left(-\Omega(\tau np)\right).
\end{aligned}$$

Moreover, by the definition of $\mathcal{A}_u$, we have

$$\mathcal{A}_u^c = (\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2)^c \cap \mathcal{F}_1.$$

Hence,

$$\begin{aligned}\mathbb{P}\left\{\mathcal{A}_u^c\right\} &= \mathbb{P}\left\{\mathcal{F}_1\right\}\mathbb{P}\left\{\left(\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2\right)^c \mid \mathcal{F}_1\right\} \\ &\le \mathbb{P}\left\{\left(\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2\right)^c \mid \mathcal{F}_1\right\} \\ &\le 8np\exp\left(-\Omega(\tau np)\right),\end{aligned}$$

completing the proof. □

The following lemma shows that with high probability, for all possible root vertex u, it has at most one children v with $|\Pi_1(v)| \le \tau$ for $\tau = o(np)$. Let A denote the adjacency matrix of G. For three distinct vertices u, v, w, define

$$\mathcal{B}_{u,v,w} = \{A_{u,v} = 1, A_{u,w} = 1\} \cap \{|\Pi_1(v)| \le \tau\} \cap \{\Pi_1(w) \le \tau\}.$$

and $\mathcal{B} = \cup_{u,v,w}\mathcal{B}_{u,v,w}$.

Lemma 16. *Assume*

$$np \ge \log n, \text{ and } np = o(n^{1/2}), \text{ and } \tau = o(np). \tag{31}$$

Then

$$\mathbb{P}\left\{\mathcal{B} \cap \mathcal{E}\right\} \le n^{-1+o(1)}.$$

Proof. By the union bound,

$$\mathbb{P}\left\{\mathcal{B} \cap \mathcal{E}\right\} \le \sum_{u,v,w} \mathbb{P}\left\{\mathcal{B}_{u,v,w} \cap \mathcal{E}\right\}$$

it reduces to bounding $\mathbb{P}\left\{\mathcal{B}_{u,v,w} \cap \mathcal{E}\right\}$.

Let N_v and N_w denote the number of neutral vertices in the branching process when v and w are popped from the head of the queue, respectively. Then conditional on N_v and N_w, $|\Pi_1(v)|$ and $|\Pi_1(w)|$ are independent and $|\Pi_1(v)| \sim \text{Binom}(N_v, p)$ and $|\Pi_1(w)| \sim \text{Binom}(N_w, p)$. On event $\mathcal{E}$, both N_v and N_w is at least $n - 1 - 4np - (4np)^2 = n - o(n)$ in view of the assumption $np = o(n^{1/2})$. Therefore,

$$\begin{aligned}&\mathbb{P}\left\{\left\{|\Pi_1(v)| \le \tau, \Pi_1(w) \le \tau\right\} \cap \mathcal{E} \mid A_{u,v} = 1, A_{u,w} = 1\right\} \\ &= \sum_{i,j=1}^{n-o(n)} \mathbb{P}\left\{\left\{|\Pi_1(v)| \le \tau, \Pi_1(w) \le \tau, N_v = i, N_w = j\right\} \cap \mathcal{E} \mid A_{u,v} = 1, A_{u,w} = 1\right\} \\ &\le \sum_{i,j=n-o(n)}^{n} \mathbb{P}\left\{|\Pi_1(v)| \le \tau, \Pi_1(w) \le \tau, N_v = i, N_w = j \mid A_{u,v} = 1, A_{u,w} = 1\right\} \\ &= \sum_{i,j=n-o(n)}^{n} \mathbb{P}\left\{N_v = i, N_w = j \mid A_{u,v} = 1, A_{u,w} = 1\right\}\mathbb{P}\left\{|\Pi_1(v)| \le \tau, \Pi_1(w) \le \tau \mid N_v = i, N_w = j\right\} \\ &= \sum_{i,j=n-o(n)}^{n} \mathbb{P}\left\{N_v = i, N_w = j \mid A_{u,v} = 1, A_{u,w} = 1\right\}\left(\mathbb{P}\left\{\text{Binom}\left(n - o(n), p\right) \le \tau\right\}\right)^2 \\ &\le \exp\left(-2(1 - o(1))np\right),\end{aligned}$$

where the last inequality holds due to the Binomial tail bound (32) and the assumption that $\tau = o(np)$. It follows that

$$\mathbb{P}\left\{\mathcal{B}_{u,v,w} \cap \mathcal{E}\right\} \le p^2\exp\left(-2(1 - o(1))np\right) = o(1/n^3),$$

where the last equality holds due to $np \ge \log n$. □

Now we are ready to prove our main proposition. Let $\mathcal{H}_u$ denote the event that tree $T(u)$ satisfies

1. u has at most one children v such that $|\Pi_1(v)| \le \tau$.
2. For each children v of u with $|\Pi_1(v)| \ge \tau$, $|\Pi_k(v)| \ge \tau \left(\frac{np}{2}\right)^{k-1}$ for all $1 \le k \le d$.

Define $\mathcal{H} = \cap_u \mathcal{H}_u$.

Proposition 1. *Suppose (28) and (31) hold and $\tau \to \infty$. Then*

$$\mathbb{P}\{\mathcal{H}\} \ge 1 - 3n^{-1+o(1)}.$$

Proof. Note that

$$(\cap_u \mathcal{A}_u) \cap (\mathcal{B}^c \cup \mathcal{E}^c) \cap \mathcal{E} \subset \mathcal{H}.$$

Hence,

$$\mathbb{P}\{\mathcal{H}\} \ge 1 - \sum_u \mathbb{P}\{\mathcal{A}_u^c\} - \mathbb{P}\{\mathcal{B} \cap \mathcal{E}\} - \mathbb{P}\{\mathcal{E}^c\}.$$

In view of Lemma 15, we have

$$\mathbb{P}\{\mathcal{A}_u^c\} \le n^{-\omega(1)}.$$

By Lemma 16, we have

$$\mathbb{P}\{\mathcal{B} \cap \mathcal{E}\} \le n^{-1+o(1)}.$$

By Lemma 1, we have $\mathbb{P}\{\mathcal{E}\} \ge 1 - 1/n$. Then the conclusion readily follows. □

Appendix E Time Complexity of Algorithm 1

Recall that in Algorithm 1, we need to efficiently check whether there exist m independent ℓ-paths from a given vertex i_2 to a set of m seeded vertices $L \subset \Gamma_\ell^{G_2}(i_2)$ in G_2 and m independent ℓ-paths from a given vertex i_1 to the corresponding seed set $\pi_0(L) \subset \Gamma_\ell^{G_1}(i_1)$ in G_1. Below we give the specific procedure to reduce this task to a maximum flow problem in a directed graph with source i_1 and sink i_2.

First, we explore the local neighborhood $N_\ell^{G_1}(i_1)$ of i_1 in G_1 up to radius ℓ. We delete all the edges (u, v) found if (u, v) are at the same distance from i_1. Also, we direct all the edges (u, v) from u to v if u is closer to i_1 than v by distance 1. Afterwards, we get a local neighborhood of i_1, denoted by $\widetilde{N}_\ell^{G_1}(i_1)$, with edges pointing away from i_1. Note that $\widetilde{N}_\ell^{G_1}(i_1)$ is not exactly a tree because a vertex may have multiple parents.

Next, we repeat the above procedure for vertex i_2 in G_2 in exactly the same manner except that the edges are directed towards i_2. Let $\widetilde{N}_\ell^{G_2}(i_2)$ denote the resulting local neighborhood of i_2.

Finally, we take the graph union of $\widetilde{N}_\ell^{G_1}(i_1)$ and $\widetilde{N}_\ell^{G_2}(i_2)$, by treating seeded vertex $u \in \Gamma_\ell^{G_2}(i_2)$ with its correspondence $\pi_0(u) \in \Gamma_\ell^{G_1}(i_1)$ as the same vertex. All the other vertices, seeded or non-seeded, from the two different local neighborhoods are treated as distinct vertices. We denote the resulting graph union as $N_\ell(i_1, i_2)$.

Recall that we aim to find independent (vertex-disjoint except for i_1) ℓ-paths from i_1 to seeded vertices in $\Gamma_\ell^{G_1}(i_1)$. Thus, we need to enforce the constraint that every vertex other than i_1 can appear at most once. Similarly for i_2. To this end, we apply the following procedure.

1. Split each vertex u in $N_\ell(i_1, i_2)$ into to two vertices: *uin* and *uout*;

2. Add an edge of capacity 1 from uin to $uout$;

3. Replace each other edge (u, v) in $N_\ell(i_1, i_2)$ with an edge from $uout$ to vin of capacity 1;

4. Find a max-flow from i_1out to i_2in.

The idea behind this construction is as follows. Any flow path from the source vertex i_1out to the sink vertex i_2in must have capacity 1, since all edges have capacity 1. Since all capacities are integral, there exists an integral max-flow in which all flows are integers [FF56]. No two flow paths can pass through the same intermediary vertex, because in passing through a vertex u in the graph the flow path must cross the edge from uin to $uout$, and the capacity here has been restricted to one. Also, the flow path must pass exactly 2ℓ distinct $uout$ vertices (including the source vertex i_1out, because all the edges are pointing away from i_1out and towards i_2in. Thus each flow path from i_1out to i_2in represents a vertex-disjoint 2ℓ-path from the source vertex i_1 to sink vertex i_2 in $N_\ell(i_1, i_2)$. As a consequence, the max-flow from i_1out to i_2in corresponds to the maximum number, m, of independent ℓ-paths from i_2 to a set of seeded vertices $L \subset \Gamma_\ell^{G_2}(i_2)$ in G_2, and of independent ℓ-paths from i_1 to the corresponding seed set $\pi_0(L) \subset \Gamma_\ell^{G_1}(i_1)$ in G_1.

As for time complexity, we can find a max-flow from i_1out to i_2in via Ford–Fulkerson algorithm [FF56] in $O(|E|f)$ time steps, where $|E|$ is the total number of edges of $N_\ell(i_1, i_2)$ after vertex splitting and edge replacement, and f is the max flow. Under the choice of ℓ given in (12), the total number of vertices and edges in $N_\ell(i_1, i_2)$ are $O(n^{1/2-\epsilon})$. Hence, $|E| = O(n^{1/2-\epsilon})$. Moreover, the max flow f is upper bounded by the number of seeded vertices in $\Gamma_\ell^{G_1}(i_1)$ which is at most $O(n^{1/2-\epsilon}\alpha)$ with high probability. Hence, in total it takes $O(n\alpha)$ time steps to compute the max-flow from i_1out to i_2in via Ford–Fulkerson algorithm.

Appendix F Tail Bounds for Binomial Distributions

Theorem 6 ([Oka59, MU05]). *Let $X \sim \mathrm{Bin}(n, p)$. It holds that*

$$\mathbb{P}\{X \le nt\} \le \exp\left(-n\left(\sqrt{p} - \sqrt{t}\right)^2\right), \quad \forall 0 \le t \le p \tag{32}$$

$$\mathbb{P}\{X \ge nt\} \le \exp\left(-2n\left(\sqrt{t} - \sqrt{p}\right)^2\right), \quad \forall p \le t \le 1. \tag{33}$$

$$\mathbb{P}\{X \ge nt\} \le 2^{-nt}, \quad \forall 6p \le t \ge 1. \tag{34}$$

Acknowledgment

J. Xu would also like to thank Cris Moore, Jian Ding, Zongming Ma, and Yihong Wu for inspiring discussions on graph matching and isomorphism. J. Xu was supported by the NSF Grant CCF-1755960.

References

[AS08] Noga Alon and Joel H. Spencer. The probabilistic method (the third edition), 2008.

[Bab16] László Babai. Graph isomorphism in quasipolynomial time [extended abstract]. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 684–697, New York, NY, USA, 2016. ACM.

[BCL+18] Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. *arXiv preprint arXiv:1805.02349*, 2018.

[BCPP98] Rainer E Burkard, Eranda Cela, Panos M Pardalos, and Leonidas S Pitsoulis. The quadratic assignment problem. In *Handbook of combinatorial optimization*, pages 1713–1809. Springer, 1998.

[BES80] László Babai, Paul Erdos, and Stanley M Selkow. Random graph isomorphism. *SIaM Journal on computing*, 9(3):628–635, 1980.

[BGM82] László Babai, D Yu Grigoryev, and David M Mount. Isomorphism of graphs with bounded eigenvalue multiplicity. In *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 310–324. ACM, 1982.

[Bol82] Béla Bollobás. Distinguishing vertices of random graphs. *North-Holland Mathematics Studies*, 62:33–49, 1982.

[Bol01] Béla Bollobás. *Random Graphs (2nd Edition)*. Cambridge Studies in Advanced Mathematics, 2001.

[CFSV04] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.

[CK16] Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for erdos-rényi graph matching. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pages 63–72. ACM, 2016.

[CK17] Daniel Cullina and Negar Kiyavash. Exact alignment recovery for correlated erdos renyi graphs. *arXiv preprint arXiv:1711.06783*, 2017.

[CP08] Tomek Czajka and Gopal Pandurangan. Improved random graph isomorphism. *Journal of Discrete Algorithms*, 6(1):85–92, 2008.

[FAP18] Donniell E. Fishkind, Sancar Adali, and Carey E. Priebe. Seeded graph matching. *arXiv preprint arXiv:1209.0367*, 2018.

[FF56] Lester R Ford and Delbert R Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8(3):399–404, 1956.

[FK15] Alan Frieze and Micha Karoski. *Introduction to Random Graphs*. Cambridge University Press, 2015.

[FQRM+16] Soheil Feizi, Gerald Quon, Mariana Recamonde-Mendoza, Muriel Médard, Manolis Kellis, and Ali Jadbabaie. Spectral alignment of networks. *arXiv preprint arXiv:1602.04181*, 2016.

[HNM05] Aria D Haghighi, Andrew Y Ng, and Christopher D Manning. Robust textual inference via graph matching. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 387–394. Association for Computational Linguistics, 2005.

[JLR11] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random Graphs*. John Wiley& Sons, Inc., 2011.

[KL14] Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*, 7(5):377–388, 2014.

[LFP13] Vince Lyzinski, Donniell E. Fishkind, and Carey E. Priebe. Seeded graph matching for correlated Erdős-Rényi graphs. *Journal of Machine Learning Research*, 15, 2013.

[Lip78] R. J. Lipton. The beacon set approach to graph isomorphism. Technical report, Yale University, 1978.

[LR13] Lorenzo Livi and Antonello Rizzi. The graph matching problem. *Pattern Analysis & Applications*, 16(3):253–283, 2013.

[Luk80] Eugene M Luks. Isomorphism of graphs of bounded valence can be tested in polynomial time. In *21st Annual Symposium on Foundations of Computer Science*, pages 42–49. IEEE, 1980.

[MMS10] Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. *Automata, Languages and Programming*, pages 594–604, 2010.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[NS09] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.

[Oka59] Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10(1):29–35, Mar 1959.

[PG11] Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1235–1243, 2011.

[SGE17] F Shirani, S Garg, and E Erkip. Seeded graph matching: Efficient algorithms and theoretical guarantees. *arXiv preprint arXiv:1805.02349*, 2017.

[SS05] Christian Schellewald and Christoph Schnörr. Probabilistic subgraph matching based on convex relaxation. In *EMMCVPR*, volume 5, pages 171–186. Springer, 2005.

[SXB08] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[Vu02] V. H. Vu. Concentration of non-lipschitz functions and applications. *Random Struct. Algorithms*, 20(3):262–316, May 2002.

[Wri71] Edward M Wright. Graphs on unlabelled nodes with a given number of edges. *Acta Mathematica*, 126(1):1–9, 1971.

[YG13] Lyudmila Yartseva and Matthias Grossglauser. On the performance of percolation graph matching. In *Proceedings of the first ACM conference on Online social networks*, pages 119–130. ACM, 2013.