Seeded graph matching via large neighborhood statistics

Elchanan Mossel | Jiaming Xu

1Department of Mathematics, MIT, Cambridge, Massachusetts
2The Fuqua School of Business, Duke University, Durham, North Carolina

Correspondence
Jiaming Xu, The Fuqua School of Business, Duke University, Durham, NC. Email: jiaming.xu868@duke.edu

Funding information
This research was supported by the National Science Foundation, IIS-1932630, CCF-1850743, and CCF-1856424.

Abstract
We study a noisy graph isomorphism problem, where 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. We show that it is possible to achieve the information-theoretic limit of graph sparsity in time polynomial in the number of vertices $n$. Moreover, we show the number of seeds needed for perfect recovery in polynomial-time can be as low as $n^\epsilon$ in the sparse graph regime (with the average degree smaller than $n^\epsilon$) and $\Omega(\log n)$ in the dense graph regime, for a small positive constant $\epsilon$. 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.

KEYWORDS
branching process; graph isomorphism; graph matching; subgraph count

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)
subexponential 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 and coworkers [3].

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 [28] and bounded eigenvalue multiplicity [6]. However, a recent breakthrough of Babai [2] gave a quasi-polynomial time algorithm.

In a number of applications including network security [31, 32], systems biology [37], computer vision [9, 36], and natural language processing [20], 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 [26], 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 $\ell_2$ 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) [4]: namely,

$$\min_{\Pi} \|G_1 - \Pi G_2 \Pi^T\|_F^2,$$

where $\Pi$ 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 [19] and the references therein). However, approximating QAP within a factor $2^{\log^{1-\epsilon}(n)}$ for $\epsilon > 0$ is NP-hard [29].

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, that is, 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 $G(n, p; s)$ [34].

**Definition 1** (The Correlated Erdős-Rényi model $G(n, p; s)$). Suppose we generate a parent graph $G_0$ from the Erdős-Rényi random graph model $G(n, p)$. 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 priori, we sample a permutation $\pi^*$ over $[n]$ uniformly at random 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 $\pi^*$ from the observation of $G_1$ and $G_2$ with high probability, that is, to design an estimator $\hat{\pi}$ based on $G_1$ and $G_2$ such that

$$\mathbb{P}\left\{ \hat{\pi}(G_1, G_2) = \pi^* \right\} \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$
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 [39] shows that if $\log n + \omega(1) \leq np \leq n - \log n - \omega(1)$, then with high probability, the size of the automorphism group of $G_0$ is $1$ and hence the underlying permutation $\pi^*$ can be exactly recovered; otherwise, with high probability, 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 [10, 11] has extended this result to the partially sampling case $s = \Theta(1)$ and $p \leq 1/2$, showing that the Maximum Likelihood Estimator, or equivalently the optimum of QAP (1), coincides with the ground truth $\pi^*$ with high probability, provided that $nps^2 \geq \log n + \omega(1)$; on the contrary, any estimator is correct with probability at most $o(1)$, if $nps^2 \leq \log n - \omega(1)$.

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 $\log n + \omega(1) \leq np \leq n - \log n - \omega(1)$ [7, 13]. However, in the partially sampling case, it is still open whether any efficient algorithm can succeed close to the threshold. A recent breakthrough result [3] obtains a quasi-polynomial-time $(n^{O(\log n)})$ algorithm which succeeds when $np \geq n^{o(1)}$ and $s \geq (\log n)^{-\omega(1)}$. However, this is still far away from the information-theoretic limit $nps^2 \geq \log n + \omega(1)$.

Another line of work [16, 22–24, 27, 34, 35, 40] 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 [31, 32]. 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 $\pi^*$, we have access to $\pi_0 \in \{1, \ldots, n\} \rightarrow \{1, \ldots, n, ?\}$ such that $\pi_0(i) = \pi^*(i)$ with probability $\alpha$ and $\pi_0(i) = ?$ with probability $1 - \alpha$ independently across different $i$, where the special symbol ? denotes the unknown vertex correspondence. The goal is to recover $\pi^*$ based on $G_1, G_2, \text{and} \pi_0$.

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

---

1In fact, a more general correlated Erdős-Rényi random graph model is considered in [10, 11], where $\mathbb{P}\{G_1(i,j) = a, G_2(i,j) = b\} = p_{a,b}$ for $a, b \in \{0, 1\}$. 
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 [25]. 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 [5] [8, Section 3.5] and empirically [24].

When \( np = \Theta(\log n) \), it is shown in [40] that if \( \alpha = \Omega(1/\log^{4/3} n) \), or equivalently, the size of the seed set is \( \Omega(n/ \log^{4/3} n) \), then a percolation-based graph matching algorithm correctly matches \( n-o(n) \) vertices in polynomial-time with high probability. The subsequent work [22] further relaxes the seed set size to be \( \Omega(n/ \log^2(n)) \) and allows for seeds to be erroneous. When \( np = n^\delta \) for some constant \( \delta \in (0,1) \), a seed set of size \( \Theta(n^{1-\delta}) \) suffices as shown in [40]. Another work [23] 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 \( n \) when \( np = \Theta(\log n) \).

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 subexponential, or polynomial time?

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

Our main results shed light on these 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 \( n^\epsilon \) 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 \( np \leq n^\epsilon \) for a fixed constant \( \epsilon < 1/6 \) and \( s = \Theta(1) \).\(^2\) Assume

\[
\begin{align*}
\text{(2)} & \quad nps^2 - \log n \to +\infty \\
\text{(3)} & \quad \alpha \geq n^{-1/2+3\epsilon}.
\end{align*}
\]

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

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

---

\(^2\)The condition \( np \leq n^\epsilon \) is only used in our analysis to ensure that condition (21) holds. Theorem 1 still holds even when \( np \geq n^\epsilon \) as long as (21) holds, for example, when \( nps^2 = n^{1/2-\epsilon} \).
Theorem 2. If
\[ nps^2 - \log n = O(1), \]
then any algorithm outputs \( \hat{\pi} \neq \pi^* \) 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 (3) 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
\[
\begin{align*}
nps^2 - \log n &\to +\infty \\
\alpha &\geq n^{-1+6\epsilon}. 
\end{align*}
\]
Then there exists a polynomial-time algorithm, namely Algorithm 3, which outputs \( \hat{\pi} = \pi^* \) with high probability 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:
\[ np = bn^a \quad (4) \]
for some fixed constants \( a, b \in (0, 1] \). Let
\[ d = \left\lceil \frac{1}{a} \right\rceil + 1. \quad (5) \]

Theorem 4. Consider the dense graph regime (4). Assume
\[ b \leq \frac{1}{16(2-s)^2}, \quad (6) \]
and
\[ \alpha \geq \frac{300 \log n}{(nps^2)^d-1}, \quad (7) \]
where \( d \) is given in (5). Then there exists an polynomial-time algorithm, namely Algorithm 2, which outputs \( \hat{\pi} = \pi^* \) with probability \( 1 - 4n^{-1} \) under the seeded \( \mathcal{G}(n, p; s, \alpha) \) model.

Note that \( d \) is the diameter of \( \mathcal{G}(n, p) \) in the regime (4) [8, Corollary 10.12]. The seed set size condition (7) stems from the fact that we match two vertices based on the number of “common” seeded vertices in their neighborhoods of radius \( d - 1 \).

Our results for seeded graph matching also imply the results for graph matching without seeds.
Theorem 5. Suppose a Seeded Graph Matching algorithm outputs $\hat{\pi} = \pi^*$ with high probability under the seeded graph matching model $\mathcal{G}(n, p; s, \alpha)$. Assume $nps^2 - \log n \to +\infty$ and $an \to +\infty$. Then there exists an algorithm, namely Algorithm 4, which calls the Seeded Graph Matching algorithm $n^{O(\alpha n)}$ times and outputs $\hat{\pi} = \pi^*$ under the seedless model $\mathcal{G}(n, p; s)$ with high probability.

Remark 1. Consider the dense regime (4) 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, $an \geq 300 \log n (bs^2)^{-k}$, 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 [3] gives a quasi-polynomial time matching algorithm in the range $np \in \left(\frac{b}{k}, \frac{n}{153}\right)$.

1.2 Key algorithmic ideas and analysis techniques

Most previous work [16, 23, 24, 27, 34, 35, 40] 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 $\ell$ for a suitably chosen $\ell$, and match two vertices by comparing the set of seeded vertices in their $\ell$th local neighborhoods. This idea was used before in the noiseless and seedless case, in [7, 13] 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 [21] and the diameter in random graphs [8]. See APPENDIX D for details.

1.3 Subsequent work on seedless graph matching

Since the initial posting of this paper to arXiv, a number of interesting papers have been posted on graph matching without the aid of initial seeds. Recent work [14] adapts the classical degree-matching algorithms in [5] and [8, Section 3.5] from the noiseless case to the noisy case, and shows that it exactly recovers $\pi^*$ with high probability in polynomial-time, provided that $p \gg \log^7(n)/n^{1/5}$ and $1 - s \ll p^4 / \log^6(n)$. Another recent work [15] develops a polynomial-time algorithm based on degree profiles (empirical distribution of the degrees of neighbors), and shows that it perfectly recovers $\pi^*$ with high probability, provided that the average degree is at least $np = \Omega(\log^2 n)$ and $1 - s = O(\log^{-2}(n))$; for dense graphs and sparse graphs, this can be further improved to $1 - s = \Omega(\log^{-2/3}(n))$ and $1 - s = O(\log^{-2}(np))$ respectively, both in polynomial time. Note that both of these two polynomial-time recovery results need the edge subsampling probability $s \to 1$ as $n \to \infty$. Finally, recent work [12]
2 | OUR ALGORITHMS

Before presenting our algorithms, we first explain why (2) is needed for graph matching under $G(n, p; s)$. Denote the intersection graph and the union graph by $G'_1 \land G_2$ and $G'_1 \lor G_2$. Then

$$G'_1 \land G_2 \sim G(n, ps^2) \quad \text{and} \quad G'_1 \lor G_2 \sim G(n, ps(2 - s)).$$

Notice that $G'_1 \land G_2$ contains the statistical signature for matching vertices, as a subgraph in $G'_1 \land G_2$ 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, $G'_1 \land G_2$ 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 $G'_1 \land G_2$ is connected with high probability. Moreover, for a high-degree vertex $i$ in $G'_1 \land G_2$, its local neighborhood grows like a branching process. In particular, the number of vertices at distance $\ell'$ from $i$ is approximately $(nps^2)^{\ell'}$. Furthermore, for a pair of two vertices $i, j$ chosen at random in $G'_1 \lor G_2$, the intersection of the local neighborhoods of $i$ and $j$ is typically of size $O((nps)^{2\ell'} n^{-1})$. Therefore, if $(nps)^{\ell'} \gg (nps)^{2\ell'} n^{-1}$ and $a(nps)^{\ell'} \gg 1$, a large number of vertices can be distinguished with high probability based on the set of seeded vertices in their $\ell'$ 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 $\Gamma^G_k(u)$ the set of vertices at distance $k$ from $u$ in $G$:

$$\Gamma^G_k(u) = \{ v \in V(G) : d(u, v) = k \} \quad (8)$$

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

$$N^G_k(u) = \bigcup_{i=0}^k \Gamma_i(u). \quad (9)$$

When the context is clear, we abbreviate $\Gamma^G_k(u)$ and $N^G_k(u)$ as $\Gamma_k(u)$ and $N_k(u)$ for simplicity. Let $d_u = |\Gamma_1(u)|$ denote the degree of vertex $u$.

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 $n^{1/2+\epsilon}$.

The idea of the algorithm is based on matching $\ell'$th local neighborhoods of two vertices by finding a set of independent paths (vertex-disjoint except for the starting vertex) to seeded vertices. The $\ell'$ 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 $a(nps^2)^{\ell'} \gg 1$, then we can find a set of sufficiently many, say $m$, independent (vertex-disjoint except for $i$) paths of length $\ell'$ from $i$ to $m$ seeded vertices in $\Gamma^G_{\ell'}(i)$. Such $m$ paths of length $\ell'$ form a starlike tree $T$ in $G'_1 \land G_2$ with root vertex $i$ and a set of $m$ seeded leaves, denoted by $L$ (See Figure 1 on page 581 for an example of $m = 3$ and $\ell' = 2$). Note that $T$
appears in $G_2$ with root vertex $i$ and the set of seeded leaves $L$; it also appears in $G_1$ with root vertex $\pi^*(i)$ and the corresponding set of seeded leaves $\pi^*(L)$. However, since the $\ell'$th local neighborhoods of two distinct vertices barely intersect, $T$ will not appear in $G^*_1 \lor G_2$ 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 a set of $m$ independent $\ell'$-paths to $m$ common seeded vertices.

**Algorithm 1** Graph matching based on counting independent $\ell'$-paths to seeded vertices

1: **Input:** $G_1, G_2, \pi_0, m, \ell' \in \mathbb{Z}$
2: **Output:** $\hat{\pi}$.
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 exists a set of $m$ independent $\ell'$-paths in $G_2$ from $i_2$ to a set of $m$ seeded vertices $L \subset \Gamma^G_{\ell'}(i_2)$, and a set of $m$ independent $\ell'$-paths in $G_1$ from $i_1$ to the corresponding set of $m$ seeded vertices $\pi_0(L) \subset \Gamma^G_{\ell'}(i_1)$, then set $\hat{\pi}(i_2) = i_1$. Declare failure if there is any conflict.
4: **Match low-degree vertices:** For every $i_2 \in I_0$, set $\hat{\pi}(i_2) = \pi_0(i_2)$. For all the pairs of unmatched vertices $(i_1, i_2)$, if $i_2$ is adjacent to a matched vertex $j_2$ in $G_2$ and $i_1$ is adjacent to vertex $\hat{\pi}(j_2)$ in $G_1$, set $\hat{\pi}(i_2) = i_1$. Declare failure if there is any conflict.
5: Output $\hat{\pi}$ to be a random permutation when failure is declared or there is any vertex unmatched.

There are two tuning parameters $\ell'$ and $m$ in Algorithm 1. Later in our analysis, we will optimally choose

$$\ell' = \left\lceil \left(\frac{1}{2} - \epsilon\right) \frac{\log n}{\log(nps^2)} \right\rceil \geq 1$$

and

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

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 $G(n, p)$ is $k$ with high probability for a fixed integer $k$, provided that

$$(k - 1)\log \log n + o(1) \leq nps^2 - \log n \leq k \log \log n - o(1),$$

see, for example, [18, 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 sets of sufficiently many 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

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

We say a vertex $i$ high-degree, if its degree $d_i \geq \tau$ in $G^*_1 \land G_2$; otherwise, we say it is a low-degree vertex. As we will see in Section 3, conditioning on that $G^*_1 \land G_2$ and $G^*_1 \lor G_2$ 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 exists a set of \(m\) independent \(\ell\)-paths in \(G_2\) from \(i_2\) to seed set \(L \subseteq \Gamma^{G_2}_{\ell}(i_2)\), and a set of \(m\) independent \(\ell\)-paths in \(G_1\) from \(i_1\) to the corresponding seed set \(\pi_0(L) \subseteq \Gamma^{G_1}_{\ell}(i_1)\). 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 [17] in \(O(na)\) 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 takes at most \(O(n^3a)\). The Step 4 of matching low-degree vertices in Algorithm 1 takes at most \(O(n^3p)\) time steps. Hence, in total Algorithm 1 takes at most \(O(n^3(a + p))\) time steps.

2.2 A simple algorithm in dense graph regime

In this subsection, we consider the dense graph regime given in (4), where \(np = bn^a\) and \(d = \lfloor 1/a \rfloor + 1\). In this setting, since \(pn^d - 2\log n \to +\infty\) and \(pn^{d-2} - 2\log n \leq -\infty\), it follows from [8, Corollary 10.12] that \(G(n, p)\) has diameter \(d\) with high probability. Thus, when \(s = \Theta(1)\), both \(G_1^s \land G_2^s\) and \(G_1^s \lor G_2^s\) 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 \(\text{within distance} (d - 1)\) from \(i_1\) in \(G_1\) and \(\text{within distance} (d - 1)\) from \(i_2\) in \(G_2\).

Algorithm 2 Graph matching based on \((d - 1)\)-hop witnesses in dense regime

1. **Input:** \(G_1, G_2, \pi_0, d \in \mathbb{Z}\).
2. **Output:** \(\hat{\kappa}\).
3. **Match all vertices:** For each pair of unseeded vertices \(i_1 \in V(G_1)\) and \(i_2 \in V(G_2)\), compute
   \[
   w_{i_1, i_2} = \left| \left\{ j \in I_0 : \pi_0(j) \in N^G_{d-1}(i_1), j \in N^G_{d-1}(i_2) \right\} \right|. \tag{13}
   \]
   Set \(\hat{\kappa}(i_2) \in \arg\max_{i_1} w_{i_1, i_2}\). Set \(\hat{\kappa}(i_2) = \pi_0(i_2)\) 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 \(N^G_{\ell}(i)\) of all vertices within distance \(\ell\) from a given vertex \(i\) takes a total of \(O(|N^G_{\ell}(i)|) = O(n)\) time steps. Thus computing \(w_{i_1, i_2}\) in (13) 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 [8, 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 \(\ell\)th local neighborhoods have a large overlap, that is,
\[ |N^G_\ell(u) \cap N^G_\ell(v)| \] is not much smaller than \(|N^G_\ell(u)|\) or \(|N^G_\ell(v)|\), rendering \(w_{i,j}\) given in (13) 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 noncommon 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 \(\ell\)th local neighborhoods (after removing vertices \(u, v\)) do not have large intersections for a suitably chosen \(\ell\). This gives rise to Algorithm 3.

Algorithm 3 Graph matching based on neighbors’ \(\ell\)-hop witnesses in sparse regime

1: **Input:** \(G_1, G_2, \pi_0, \ell \in \mathbb{Z}, \eta \in \mathbb{R}_+\).

2: **Output:** \(\hat{\pi}\).

3: **Match high-degree vertices:** For all the pairs of unseeded vertices \((u, v)\) and for each pair of their neighbors \((i, j)\) with \(i \in \Gamma^G_1(u)\) and \(j \in \Gamma^G_1(v)\), compute

\[
w_{i,j}^{u,v} = \min_{x \in V(G_1), y \in V(G_2)} \left\{ \left\{ k \in I_0 : \pi_0(k) \in N^G_\ell(x \setminus \{i\}) \land k \in N^G_\ell(y \setminus \{j\}) \right\} \right\}, \tag{14}
\]

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

\[
Z_{u,v} = \sum_{i \in \Gamma^G_1(u)} \sum_{j \in \Gamma^G_1(v)} 1 \{ w_{i,j}^{u,v} \geq \eta \}. \tag{15}
\]

If \(Z_{u,v} \geq \log n / \log \log n - 1\), set \(\hat{\pi}(v) = u\). 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 \(\ell\) from both vertex \(i\) in \(G_1\) and vertex \(j\) in \(G_2\) in (14), 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,

\[
w_{i,j}^{u,v} \leq \left\{ k \in I_0 : \pi_0(k) \in N^G_\ell(x \setminus \{i\}) \land k \in N^G_\ell(y \setminus \{j\}) \right\}, \tag{16}
\]

where the right hand side becomes independent from the edges incident to \(u\) and \(v\) in \(G_1^* \lor G_2\). 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 [38].

There are two tuning parameters \(\ell\) and \(\eta\) in Algorithm 3. In our analysis later, we will optimally choose

\[
\ell = \left\lfloor \frac{(1 - \epsilon) \log n}{\log(np)} \right\rfloor, \tag{17}
\]

and

\[
\eta = 4^{2\ell + 2} p^{2\ell} n^{2\ell - 1} \alpha. \tag{18}
\]

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 (16) for given pairs of vertices $(u, v)$ and $(i, j)$. There are at most $n^{2+2c}$ such pairs. The step of matching low-degree vertices as specified in Algorithm 1 takes $O(n^3p)$ time steps in total. Thus in total Algorithm 3 takes at most $O(n^{5+2c})$ 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 $f : I_0 \rightarrow [n]$ 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:** $\hat{\Pi}$.
3: Select a random subset $I_0$ of $V(G_1)$ by including each vertex with probability $\alpha$.
4: For every possible mapping $f : I_0 \rightarrow [n]$, run Seeded Graph Matching Algorithm with a seed set $I_0$, and output $\pi_f$.
5: Output $\hat{\pi} \in \arg \min_{\pi_f} \|G_1 - \Pi_f G_2 \Pi_f^T\|^2_{F}$, where $\Pi_f$ is the permutation matrix corresponding to $\pi_f$.

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 $f : I_0 \rightarrow [n]$ 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 [11, Theorem 1] that with high probability,

$$\pi^* \in \arg \min_{\pi} \|G_1 - \Pi G_2 \Pi^T\|^2_{F},$$

where $\Pi$ is the permutation matrix corresponding to $\pi$.

Therefore, $\hat{\pi} = \pi^*$ with high probability. Finally, since $an \rightarrow \infty$, it follows that $|I_0|$ is at most $2an$ with high probability. Hence, Algorithm 4 calls the Seeded Graph Matching algorithm at most $n^{O(an)}$ 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 $\pi^* = id$, that is, $\pi^*(i) = i$ for all $i \in [n]$, without loss of generality.
Success of Algorithm 1 on the intersection of good events

3. For all vertices $i$

2. For any two adjacent vertices, there are at least $\tau$ vertices adjacent to at least one of them;
3. For all vertices $i$ with $d_i \geq \tau$, there exists a set of $2m$ independent $\ell'$-paths from $i$ to $2m$ distinct vertices in $I_\ell$;
4. There is no pair 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 Figure 1 on page 581 for an illustration).
5. For every vertex $i$, there is no set of $m$ independent $\ell'$-paths from $i$ to $m$ distinct vertices in $N^G_{\ell-1}(i)$.

Let $\ell$, $m$, and $\tau$ be given in (10), (11), and (12), respectively. Let

- $E_1$ denote the event such that $G_1 \land G_2$ satisfy properties (1)-(3);
- $E_2$ denote the event such that $G_1 \lor G_2$ satisfy properties (4) and (5);
- $E_3$ denote the event such that for any two vertices $i, j$ that are connected by a 2-path in $G_1 \lor G_2$, at least one of the two vertices $i, j$ must be a high-degree (of degree at least $\tau$) vertex in $G_1 \land G_2$.

We claim that on event $E_1 \cap E_2 \cap E_3$, Algorithm 1 correctly matches all vertices.
First, since $G_1 \land G_2$ satisfies graph property (iii), it follows that in $G_1 \land G_2$, for all high-degree vertices $i$ (of degree at least $\tau$), there exists a set of $2m$ independent $\ell$-paths to a set $S \subset N_{\ell}^{G_1 \land G_2}(i)$ of $2m$ distinct seeded vertices. Let $\tilde{S} = S \setminus N_{\ell-1}^{G_1 \land G_2}(i)$. Since $G_1 \lor G_2$ satisfies graph property (v) and $G_1 \land G_2 \subset G_1 \lor G_2$, it follows that

$$|S \cap N_{\ell-1}^{G_1 \lor G_2}(i)| \leq m - 1$$

and thus $|\tilde{S}| \geq m + 1$. Moreover, since $G_1 \land G_2 \subset G_1, G_2 \subset G_1 \lor G_2$, it follows that

$$\tilde{S} \subset N_{\ell}^{G_1 \land G_2}(i) \setminus N_{\ell-1}^{G_1 \lor G_2}(i) \subset \Gamma_{\ell}^{G_1}(i) \cap \Gamma_{\ell}^{G_2}(i).$$

Therefore, in both $G_1$ and $G_2$, there exists a set of $m + 1$ independent $\ell$-paths from $i$ to $m + 1$ distinct seeded vertices in $\Gamma_{\ell}^{G_1}(i) \cap \Gamma_{\ell}^{G_2}(i)$.

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

It follows from the above two points that Algorithm 1 correctly matches all high-degree vertices $i$ in $G_1 \land G_2$, that is, $\hat{\pi}(i) = i$.

Next, we show that all low-degree vertices are matched correctly. Fix a low-degree vertex $i$. Since $G_1 \land G_2$ satisfies graph properties (i) and (ii), it must have a high-degree neighbor $j$ in $G_1 \land G_2$. Since the high-degree vertex $j$ has been matched correctly, $i$ is adjacent to $j$ in $G_1$ and $i$ is also adjacent to $j = \hat{\pi}(j)$ 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 $\hat{\pi}(j_1)$ in $G_2$. Since $j_1$ is matched correctly, i.e., $\hat{\pi}(j_1) = j_1$, it follows that $(i_1, j_1, i_2)$ form a 2-path in $G_1 \lor G_2$. However, on event $\mathcal{E}_3$, $i_1$ and $i_2$ cannot be low-degree vertices simultaneously in $G_1 \land G_2$, which leads to a contradiction. As a consequence, Algorithm 1 correctly matches all low-degree vertices $i$ in $G_1 \land G_2$, that is, $\hat{\pi}(i) = i$.

Finally, to prove Theorem 1, it remains to show that under the theorem assumptions, $\mathbb{P}\{\mathcal{E}_1\} \to 1$ 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 $G_1 \land G_2$ 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 high probability;

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

**Lemma 2.** Assume

$$np^2 s \geq \log n \quad \text{and} \quad \tau = o(np^2) \quad \text{and} \quad \log(np) = o(np^2).$$
With probability at least $1 - n^{-1+o(1)}$, for any two vertices $i, j$ that are connected by a 2-path in $G_1 \lor G_2$, at least one of the two vertices $i, j$ must have degree at least $\tau$ in $G_1 \land G_2$.

It remains to show with high probability, $G_1 \land G_2$ satisfies graph property (iii) and $G_1 \lor G_2$ satisfies graph properties (iv) and (v).

We will apply the following lemma to show that with high probability, for every high-degree vertex in $G_1 \land G_2$, there exists a set of $2m$ independent paths of length $\ell'$ from $i$ to $2m$ distinct seeded vertices in $I_0$.

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

$$
\log n \leq np = o\left(n^{1/2}\right), \quad \text{and} \quad o(1) = \tau = o(np).
$$

Let $m, \ell \in \mathbb{N}$ be two positive integers such that

$$
\alpha(np/2)^{\ell-2} \tau(\tau - 2m) - 2m \log \tau \geq 2 \log n, \quad \text{and} \quad p(4np)^\ell = o(1).
$$

Then with probability at least $1 - 5n^{-1+o(1)}$, for all vertices $i$ with $d_i \geq \tau$, there exists a set of at least $2m$ independent $\ell'$-paths from $i$ to $I_0$.

**Proof.** In APPENDIX D.3, we describe a graph branching process to construct tree $T_{\ell'}(i) \subset G$ of depth $\ell'$ rooted at $i$ for every vertex $i$. Let $H$ denote the event that for every vertex $i$, $T_{\ell'}(i)$ satisfies

1. Root $i$ has $d_i$ children and at most one child $j$ has fewer than $\tau$ children in $T_{\ell'}(i)$, that is, $|\Pi_1(j)| < \tau$;
2. For each child $j$ 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, that is, $|\Pi_{\ell'-1}(j)| \geq \tau(np/2)^{\ell'-2}$,

where $\Pi_k(j)$ for $1 \leq k \leq \ell' - 1$ denotes the set of vertices at distance $k$ from $j$ in the subtree $T_{\ell'-1}(j)$. We show that $\mathbb{P}\{H\} \geq 1 - 3n^{-1+o(1)}$ by Proposition 1 under the lemma assumptions.

Condition on $G$ such that event $H$ holds. Fix a vertex $i$ and tree $T_{\ell'}(i)$. Then $i$ has at least $d_i - 1$ children $j$ such that $|\Pi_1(j)| \geq \tau$. For each such $j$, define $Y_{ij} = 0$ if $I_0 \cap \Pi_{\ell'-1}(j) = \emptyset$; and $Y_{ij} = 1$ otherwise. For each such $j$ such that $Y_{ij} = 1$, choose a vertex in $I_0 \cap \Pi_{\ell'-1}(j)$ and include the unique path in $T_{\ell'}(i)$ from $i$ to this vertex into a set $S_i$. By construction, $S_i$ consists of independent $\ell'$-paths from $i$ to $I_0$ and $|S_i| \geq \sum_{j=1}^{d_i-1} Y_{ij}$. It remains to lower bound $\sum_{j=1}^{d_i-1} Y_{ij}$.

Since each vertex in $\Pi_{\ell'-1}(j)$ is included into $I_0$ with probability $\alpha$ independently across different vertices and from graph $G$, it follows that conditional on $G$, $Y_{ij}$ are mutually independent across different $j$ and

$$
\mathbb{P}\{Y_{ij} = 0 \mid G\} 1_{\{H\}} = (1 - \alpha)^{|\Pi_{\ell'-1}(j)|} 1_{\{H\}} \leq \exp\left(-\alpha \tau(np/2)^{\ell'-2}\right),
$$

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

$$
\mathbb{P}\left\{\sum_{j=1}^{d_i-1} Y_{ij} \leq 2m - 1 \mid G\right\} 1_{\{(d_i, \tau) \cap H\}} \leq \mathbb{P}\left\{\sum_{j=1}^{\tau-1} Y_{ij} \leq 2m - 1 \mid G\right\} 1_{\{(d_i, \tau) \cap H\}} \leq \mathbb{P}\left\{\text{Binom}\left(\tau - 1, 1 - e^{-\alpha \tau(np/2)^{\ell'-2}}\right) \leq 2m - 1\right\},
$$
\[
\begin{align*}
&\leq \sum_{k=0}^{2m-1} \binom{\tau - 1}{k} e^{-a\tau(np/2)^{\ell-2}(\tau-1-k)} \\
&\leq e^{-a\tau(np/2)^{\ell-2}(\tau-2m)} \sum_{k=0}^{2m-1} \tau^k \\
&\leq 2e^{-a\tau(np/2)^{\ell-2}(\tau-2m)} \tau^{2m} \leq 2n^{-2},
\end{align*}
\]

where the last inequality holds due to the assumption \(a(np/2)^{\ell-2}\tau(\tau - 2m) - 2m \log \tau \geq 2 \log n\).

Define event
\[
P_i = \{d_i \geq \tau\} \cap \left\{\sum_{j=1}^{d_i-1} Y_{ij} \leq 2m - 1\right\}.
\]

Then we have that
\[
P \{F_i \cap H\} = \mathbb{E}_G \left[\mathbb{E} \left[\mathbf{1}\left\{\sum_{j=1}^{d_i-1} Y_{ij} \leq 2m - 1\right\} \mathbf{1}\{\{d_i, \geq \tau\} \cap H\} \mid G\right]\right]
\]
\[
= \mathbb{E}_G \left[\mathbb{P} \left\{\sum_{j=1}^{d_i-1} Y_{ij} \leq 2m - 1 \mid G\right\} \mathbf{1}\{\{d_i, \geq \tau\} \cap H\}\right]
\]
\[
\leq 2n^{-2}.
\]

Let \(F = \bigcup_i F_i\). By the union bound,
\[
P \{F\} \leq P \{F \cap H\} + P \{H^c\} \leq \sum_i P \{F_i \cap H\} + P \{H^c\} \leq 2n^{-1} + 3n^{-1+o(1)} \leq 5n^{-1+o(1)}.
\]

Therefore, with probability at least \(1 - 5n^{-1+o(1)}\), for all vertices \(i\) with \(d_i \geq \tau\), \(\sum_{j=1}^{d_i-1} Y_{ij} \geq 2m\) and hence \(S_i\) consists of at least \(2m\) independent \(\ell\)-paths from \(i\) to \(I_0\).

The following lemma will be useful to conclude that in \(G_1 \lor G_2\), with high probability, there is no pair of subgraphs \(T_1, T_2 \subset G_1 \lor G_2\) that are isomorphic to \(T\), such that \(r(T_1) \neq r(T_2)\) and \(L(T_1) = L(T_2)\).

See Figure 1 on page 581 for an illustration of \(T_1\) and \(T_2\) isomorphic to \(T\) such that \(r(T_1) \neq r(T_2)\) and \(L(T_1) = L(T_2)\).

**Lemma 4.** Suppose \(G \sim G(n, p)\) and \(\ell, m \geq 1\). Then it holds that
\[
P \left\{\exists T_1, T_2 \subset G \text{ that are isomorphic to } T : r(T_1) \neq r(T_2), L(T_1) = L(T_2)\right\}
\]
\[
\leq \left(2 + \frac{8}{np}\right)^{m(\ell-1)} n^{2m\ell+2-m} p^{2m\ell}.
\]

**Proof.** Let \(T\) denote the set of all possible subgraphs that are isomorphic to \(T\) in the complete graph \(K_n\). By the union bound, we have
\[
P \left\{\exists T_1, T_2 \subset G \text{ that are isomorphic to } T : r(T_1) \neq r(T_2), L(T_1) = L(T_2)\right\}
\]
\[
\leq \sum_{T_1, T_2 \in T : r(T_1) \neq r(T_2), L(T_1) = L(T_2)} P \{T_1, T_2 \subset G\}.
\]
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 $|E(T)| = 2m\ell$.

For any given unlabelled graph $S$, let $n_S$ denote the 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 $L(T_1) = L(T_2)$. Therefore,

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

Next we upper bound $n_S$. Let $\kappa_S$ denote the number of subgraphs $S'$ in $T$ such that $S'$ is isomorphic to $S$, $L(T) \subset V(S')$, and $r(T) \not\in V(S')$. Then there are at most $\kappa_S^2$ 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 $(T_1 \setminus T_2) \cup (T_2 \setminus T_1)$. Hence,

$$n_S \leq \kappa_S^2 n'|V(S')| n^2(|V(T)| - |V(S)|) = \kappa_S^2 n^{2m\ell + 2 - |V(S)|},$$

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 $\kappa_S$, $S$ is isomorphic to some $S' \subset T$ such that $L(T) \subset V(S')$ and $r(T) \not\in V(S')$. By the starlike tree property of $T$, $S'$ is a forest with at least $m$ disjoint trees; hence so is $S$. See Figure 1 on page 581 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

$$\sum_{S : |V(S)| = m + k} \kappa_S \leq \binom{m(\ell - 1)}{k} 2^k,$$

because there are at most $\binom{m(\ell - 1)}{k}$ different choices for $V(S') \setminus L(T)$ and at most $2^{|V(S')| - m}$ choices for determining whether to include the edges induced by $V(S')$ in $T$ into $S'$. Hence,

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

where (a) follows from \( \sum_{|S|:|V(S)|=m+k} k_S \leq \binom{m(\ell-1)}{k} 2^k \), and (b) holds due to \( \binom{m(\ell-1)}{k} \leq 2^{m(\ell-1)} \).}

Finally, we need a result to conclude that with high probability, for every vertex \( i \), there is no set of \( m \) independent \( \ell \)-paths from \( i \) to \( m \) distinct vertices in \( N_{\ell-1} G_1 \cup G_2(i) \).

Fix \( m, \ell \geq 1 \). We start with vertex \( i \) and a set of \( m \) independent (vertex-disjoint except for \( i \)) paths of length \( \ell \) from \( i \) to \( m \) distinct vertices \( j_1, \ldots, j_m \), denoted by \( P_1, \ldots, P_m \). Let \( \tilde{P}_k \) denote a path of length at most \( \ell - 1 \) from \( i \) to \( j_k \) for \( k = 1, \ldots, m \). Let \( H = \cup_{k=1}^m (P_k \cup \tilde{P}_k) \). Define \( \mathcal{H}_{m, \ell} \) to be 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 \( G_1 \cup G_2 \), then for every vertex \( i \), there is no set of \( m \) independent \( \ell \)-paths from \( i \) to \( m \) distinct vertices in \( N_{\ell-1} G_1 \cup G_2(i) \). Hence, our task reduces to proving that with high probability, \( G_1^* \cup G_2 \) 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 is unlikely to appear as a subgraph in \( G(n, p) \).

**Lemma 5.** Fix \( m, \ell \geq 1 \). For any \( H \in \mathcal{H}_{m, \ell} \),

\[ |E(H)| \geq |V(H)| + m - 1. \]

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

We order the vertices and edges in paths starting from \( i \). For each \( k = 1, \ldots, m \), let \( v_k \) denote the first vertex starting from which \( i \) is on the \( \ell \)-path from \( i \) to \( j_k \), then by definition, \( v_k \neq i \). Let \( \text{dist}(u, v) \) denote the shortest distance between \( u \) and \( v \) in \( H \), and \( \sigma \) denote any permutation on \([m]\) such that

\[ \text{dist} \left( i, v_{\sigma(1)} \right) \geq \text{dist} \left( i, v_{\sigma(2)} \right) \cdots \geq \text{dist} \left( i, v_{\sigma(m)} \right). \]

Without loss of generality, we assume \( \sigma = \text{id} \), that is, \( \sigma(k) = k \). Let \( e_k \) denote the edge before \( v_k \) in \( \tilde{P}_k \), that is, \( e_k \) is the edge incident to \( v_k \) in \( \tilde{P}_k \) but not \( P_k \). Then by definition \( e_k \notin P_k \). Let \( Q_k \) denote the subpath of \( P_k \) from \( i \) to \( v_k \), and \( \tilde{Q}_k \) denote the subpath of \( \tilde{P}_k \) from \( i \) to \( v_k \).

First, we claim that there is a cycle containing \( e_k \) in \( Q_k \cup \tilde{Q}_k \). This is because both \( Q_k \) and \( \tilde{Q}_k \) are paths from \( i \) to \( v_k \), and \( e_k \in \tilde{Q}_k \setminus Q_k \).

Next, we claim that \( e_j \notin \tilde{Q}_k \cup \tilde{Q}_k \) for any \( 1 \leq j < k \leq m \). In fact, \( e_j \notin \tilde{Q}_k \), because otherwise \( P_j \) and \( P_k \) share a common vertex \( v_j \neq i \), which violates the assumption that \( P_j \) and \( P_k \) are vertex-disjoint except for \( i \). Also, \( e_j \notin \tilde{Q}_k \). Suppose not. Then \( v_j \) is on the path \( \tilde{Q}_k \). Since \( P_j \) and \( P_k \) are vertex-disjoint except for \( i \), it follows that \( v_j \neq v_k \). Therefore, \( v_j \) is ordered strictly before \( v_k \) in the path \( \tilde{Q}_k \) starting from \( i \). Since \( \tilde{P}_k \) is a shortest path from \( i \) to \( j_k \), by the optimal substructure property, \( \tilde{Q}_k \) is a shortest path from \( i \) to \( v_k \). It follows that \( \text{dist}(i, v_j) < \text{dist}(i, v_k) \), which leads to a contradiction.
Finally, we recursively define $H_0 = H$ and $H_k$ such that $V(H_k) = V(H_{k-1})$ and $E(H_k) = E(H_{k-1}) \setminus \{e_k\}$ 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 $e_j \notin Q_k \cup \tilde{Q}_k$ for any $1 \leq j < k \leq m$, it follows that $Q_k \cup \tilde{Q}_k \subset H_{k-1}$. Also, we have shown that there is a cycle containing $e_k$ in $Q_k \cup \tilde{Q}_k$. 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.

It follows from induction that $H_m$ is connected. Thus, $|E(H_m)| - |V(H_m)| \geq -1$. Since $|E(H)| = |E(H_m)| + m$ and $|V(H)| = |V(H_m)|$, it follows that $|E(H)| - |V(H)| \geq m - 1$. 

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

**Lemma 6.** Fix $m, \ell \geq 1$. Denote by $U_{m, \ell}$ the set of unlabelled graphs (isomorphism classes) in $H_{m, \ell}$. Then

$$|U_{m, \ell}| \leq (3\ell)^m 3^{2m^2 \ell}.$$

**Proof.** Recall that $H = \bigcup_{k=1}^{m}(P_k \cup \tilde{P}_k)$, where $P_1, \ldots, P_m$ is a set of $m$ (vertex-disjoint except for $i$) paths of length $\ell$ from $i$ to $m$ distinct vertices $j_1, \ldots, j_m$, and $\tilde{P}_k$ is a path of length at most $\ell - 1$ from $i$ to $j_k$ for $k = 1, \ldots, m$. Let $T = \bigcup_{k=1}^{m}P_k$. Then $T$ is a starlike tree rooted at $i$ with $m$ branches as depicted in Figure 1 on page 581.

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

Then across all $1 \leq j < k \in [m]$, there are at most $\kappa_{\ell}(S)\kappa_{\ell}(S)$ ways of intersecting $\tilde{P}_j$ and $\tilde{P}_k$ such that $\tilde{P}_j \cap \tilde{P}_k$ is isomorphic to $S$. Also, for all $k \in [m]$, there are at most $\beta(S)\kappa_{\ell}(S)$ ways of intersecting $\tilde{P}_0$ and $\tilde{P}_k$ such that $\tilde{P}_0 \cap \tilde{P}_k$ is isomorphic to $S$. Hence, the total number of distinct sequences of $(\tilde{P}_1, \ldots, \tilde{P}_k)$ such that $\tilde{P}_j \cap \tilde{P}_k$ is isomorphic to $S_{jk}$ for $0 \leq j < k \leq m$ is at most

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

Therefore,

$$|U_{\ell_1, \ldots, \ell_m}| \leq \sum_{\{S_{jk} : 0 \leq j < k \leq m\}} \prod_{j \leq k \in [m]} \kappa_{\ell}(S_{jk})\kappa_{\ell}(S_{jk}) \prod_{k \in [m]} \beta(S_{0k})\kappa_{\ell}(S_{0k})$$

$$\leq \prod_{1 \leq j < k \in [m]} \left(\sum_{S_{jk}} \kappa_{\ell}(S_{jk})\right)^2 \beta(S_{0k})\kappa_{\ell}(S_{0k}) \prod_{k \in [m]} \left(\sum_{S_{0k}} \beta(S_{0k})\right) \left(\sum_{S_{jk}} \kappa_{\ell}(S_{jk})\right)\left(\sum_{S_{0k}} \kappa_{\ell}(S_{0k})\right)$$

$$\leq \prod_{j \leq k \in [m]} \frac{n_{\ell}}{n_{\ell}} \prod_{k \in [m]} \left(n(T)n_{\ell}\right)^m \prod_{k \in [m]} \left(n_{\ell}\right)^m.$$
where the last inequality holds because \( \sum_S \kappa(S) \) is at most the the number \( n_{\ell} \) of distinct subgraphs \( S' \) in an \( \ell' \)-path, and \( \sum_S \beta(S) \) is at most the the number \( n(T) \) of distinct subgraphs \( S' \) in \( T \). Note that

\[
    n_{\ell} \leq \sum_{k=0}^{\ell} \binom{\ell}{k} 2^k = 3^{\ell},
\]

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 \( \ell' \)-path into \( S' \). Analogously,

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

Combining the last three displayed equations yields that

\[
    \left| \mathcal{U}_{\ell_1,\ldots,\ell_m} \right| \leq 3^{2m\ell+m}.
\]

Therefore,

\[
    \left| \mathcal{U} \right| = \sum_{(\ell_1,\ldots,\ell_m):1 \leq \ell_1 \leq \ell' - 1} \left| \mathcal{U}_{\ell_1,\ldots,\ell_m} \right| \leq (3\ell')^m 3^{2m\ell}.
\]

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

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

\[
    \mathbb{P} \left\{ \exists H \in \mathcal{H}_{m,\ell} : H \subset G \right\} \leq n^{2m\ell-2m+1} p^{2m\ell-m} (3\ell')^m 3^{2m\ell}. \tag{19}
\]

**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 \( \tilde{P}_k \) 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

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

and let \( \mathcal{U}_{m,\ell,t} \) denote the number of unlabelled graphs (isomorphism class) in \( \mathcal{H}_{m,\ell,t} \). Since \( V(H) \subset [n] \) and \( |V(H)| = t \), there are at most \( n' \) different vertex labelings for a given unlabelled graph \( U \in \mathcal{U}_{m,\ell,t} \). Hence,

\[
    \left| \mathcal{H}_{m,\ell,t} \right| \leq \left| \mathcal{U}_{m,\ell,t} \right| n'. \tag{20}
\]

By the union bound, we have

\[
    \mathbb{P} \left\{ \exists H \in \mathcal{H}_{m,\ell} : H \subset G \right\} \leq \sum_{t=m\ell+1}^{2m\ell-2m+1} \sum_{H \in \mathcal{H}_{m,\ell,t}} \mathbb{P} \{ H \subset G \}.
\]
Recall that the choices of \( \ell \) in (10), \( m \) in (11), and \( \tau \) in (12). In particular, under the assumption that \( np \leq n^\epsilon \),
\[
n^{1/2-2\epsilon} \leq (nps^2)^\ell \leq n^{1/2-\epsilon}.
\]

Recall that \( G_1 \land G_2 \sim \mathcal{G}(n, ps^2) \). Under the assumption that \( a \geq n^{-1/2+3\epsilon} \), we get that for \( n \) sufficiently large,
\[
a(nps^2/2)^\ell^{-2} \tau(\tau - m) - m \log \tau \geq 2 \log n.
\]
Hence, applying Lemma 3, we conclude that \( G_1 \land G_2 \) satisfies graph property (iii). Combining this result with Lemma 1, we get that \( \mathbb{P} \{ E_1 \} \geq 1 - o(1) \).

Note that \( G_1 \lor G_2 \sim \mathcal{G}(n, ps(2-s)) \). We first apply Lemma 4 to \( G_1 \lor G_2 \). 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
\[
\left( 2 + \frac{8}{nps(2-s)} \right)^{m(\ell-1)} n^{2m\ell+2m} (ps(2-s))^{2m\ell} \leq 10^{m\ell} (nps^2)^{2m\ell} n^{2m} \left( \frac{2-s}{s} \right)^{2m\ell} \leq 10^{m\ell} n^{2-2e}m \left( \frac{2-s}{s} \right)^{2m\ell} = n^{-2+o(1)},
\]
where the second inequality holds due to \( (nps^2)^\ell \leq n^{1/2-\epsilon} \); the last equality holds by our choice of \( \ell \) and \( m \) and \( s = \Theta(1) \). Hence, applying Lemma 4 to \( G_1 \lor G_2 \), we conclude that with high probability, there is no pair of subgraphs \( T_1, T_2 \subseteq G_1 \lor G_2 \) 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 \( G_1 \lor G_2 \). Note that
\[
n^{2m\ell-2m+1} (ps(2-s))^{2m\ell-m} (3\epsilon)^m 32m^2\ell
\]
\[
= (nps^2)^{2m\ell} n^{m-1} (np)^{-m} \left( \frac{2-s}{s} \right)^{2m\ell-m} (3\epsilon)^m 32m^2\ell
\]
Recall that $\Gamma_N^m$ and $d$ where the last inequality holds due to our choice of $\ell'$ and $m$ and $s = \Theta(1)$. Hence, applying Lemma 7 to $G_1 \cup G_2$, we conclude that with high probability, $G_1 \cup G_2$ 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 is no set of $m$ independent paths from $i$ to $m$ distinct vertices in $\Gamma_{\ell - 1}^{G_1 \cup G_2}$.

Combining the above two points, we get that $\mathbb{P}\{E_2\} \to 1$. Finally, in view of Lemma 2, we get that $\mathbb{P}\{E_3\} \geq 1 - o(1)$, completing the proof of Theorem 1.

4 | ANALYSIS OF ALGORITHM 2 IN DENSE GRAPH REGIME

Recall that $\Gamma_k^G(u)$ and $N_k^G(u)$ denote the set of vertices at and within distance $k$ from $u$ in graph $G$, as defined in (8) and (9), respectively. The key is to show that $|N_{d-1}^{G_1 \cup G_2}(u)|$ is larger than $|N_{d-1}^{G_1 \cup G_2}(u) \cap N_{d-1}^{G_1 \cup G_2}(v)|$ for $u \neq v$, so that we can match two vertices correctly based on the number of common seeded vertices in their two large neighborhoods.

**Proof of Theorem 4.** Define event

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

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

Define event

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

Note that in the dense graph regime (4) with $np = bn^a$,

$$\frac{(nps^2)^{d-1}}{16n^{d-3}(ps(2-s))^{2d-2}} = \frac{n}{16(np)^{d-1}(2-s)^{2d-2}} = \frac{n}{16n^{a(d-1)}b^{d-1}(2-s)^{2d-2}} \geq \frac{16^{d-2}n}{n^{a(d-1)}}.$$  

where the last inequality holds due to the assumption (6). Since $d - 1 = \lfloor 1/a \rfloor$, it follows that for $n \geq 16$ and all $d \geq 1$,

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

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

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

$$C = \left\{ N_{d-1}^{G_1 \land G_2} \cap I_0 > \frac{3}{2}(nps^2)^{d-1}a, \ \forall u \right\}.$$  

Since each vertex is seeded independently with probability $\alpha$, it follows that

$$\mathbb{P}\{C^c\} \leq \mathbb{P}\{A^c\} + \mathbb{P}\{C^c \mid A\}.$$
\[
\leq n^{-10} + \sum_u \mathbb{P} \left\{ \left| N_{d-1}^{G_1 \land G_2}(u) \cap I_0 \right| \leq \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha \right\} 
\]
\[
\leq n^{-10} + n \mathbb{P} \left\{ \text{Bin} \left( \frac{3}{4} \left( nps^2 \right)^{d-1} \right), \alpha \right\} \leq \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha \right\} 
\]
\[
\leq n^{-10} + n \exp \left( -\frac{3}{200} \left( nps^2 \right)^{d-1} \alpha \right) \leq 2n^{-1},
\]

where the second to the last inequality holds because \( \mathbb{P} \{ \text{Bin}(n, p) \leq (1 - \epsilon)np \} \leq e^{-\epsilon^2 np/2} \) for \( \epsilon \in [0, 1] \); the last inequality holds due to assumption (7).

Similarly, define event
\[
D = \left\{ \left| N_{d-1}^{G_1 \land G_2}(u) \cap N_{d-1}^{G_1 \land G_2}(v) \cap I_0 \right| < \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha, \forall u \neq v \right\}.
\]

It follows that
\[
\mathbb{P} \{ D^c \} \leq \mathbb{P} \{ B^c \} + \mathbb{P} \{ D^c \mid B \}
\]
\[
\leq n^{-10} + \sum_u \mathbb{P} \left\{ \left| N_{d-1}^{G_1 \lor G_2}(u) \cap N_{d-1}^{G_1 \lor G_2}(v) \cap I_0 \right| \geq \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha \mid B \right\}
\]
\[
\leq n^{-10} + n \mathbb{P} \left\{ \text{Bin} \left( \frac{1}{2} \left( nps^2 \right)^{d-1} \right), \alpha \right\} \leq \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha \right\}
\]
\[
\leq n^{-10} + n \exp \left( -\frac{1}{150} \left( nps^2 \right)^{d-1} \alpha \right) \leq 2n^{-1},
\]

where the second to the last inequality holds because \( \mathbb{P} \{ \text{Bin}(n, p) \geq (1 + \epsilon)np \} \leq e^{-\epsilon^2 np/3} \) for \( \epsilon \in [0, 1] \); the last inequality holds again due to assumption (7). Hence, \( \mathbb{P} \{ C \cap D \} \geq 1 - 4n^{-1} \).

Finally, since \( G_1 \land G_2 \) is a subgraph of both \( G_1 \) and \( G_2 \), it follows that
\[
N_{d-1}^{G_1 \land G_2}(i_2) \cap I_0 \subseteq \left\{ j \in I_0 : \pi_0(j) \in N_{d-1}^{G_1}(\pi^*(i_2)), j \in N_{d-1}^{G_2}(i_2) \right\}.
\]

Similarly, both \( G_1 \) and \( G_2 \) are subgraphs of \( G_1 \lor G_2 \), it follows that
\[
\left\{ j \in I_0 : \pi_0(j) \in N_{d-1}^{G_1}(i_1), j \in N_{d-1}^{G_2}(i_2) \right\} \subseteq N_{d-1}^{G_1 \lor G_2} \left( (\pi^*)^{-1}(i_1) \right) \cap N_{d-1}^{G_1 \lor G_2}(i_2) \cap I_0.
\]

Thus recalling the definition of \( w_{i_1, i_2} \), we get that on event \( C \cap D \), for every vertex \( i_2 \in V(G_2) \setminus I_0 \),
\[
w_{i_1, i_2} \begin{cases} 
> \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha & \text{if } i_1 = \pi^*(i_2), \\
< \frac{3}{5} \left( nps^2 \right)^{d-1} \alpha & \text{otherwise}.
\end{cases}
\]

Hence, Algorithm 2 outputs \( \hat{\pi} = \pi^* \) on event \( C \cap D \).

5 | ANALYSIS OF ALGORITHM 3 IN SPARSE GRAPH REGIME

Before proving Theorem 3, we present two key lemmas. The first lemma will be used later to conclude that the test statistic \( Z_{\alpha,u} \) given in (15) is large for all high-degree vertices \( u \).

**Lemma 8.** Suppose \( G \sim G(n, p) \) with \( \log n \leq np \leq n^{\epsilon} \), and each vertex is included in \( I_0 \) with probability \( \alpha \). Recall that \( \epsilon \) and \( \eta \) are given in (17) and (18), respectively. Assume \( \eta \geq 6 \log n. \)
Let $G \setminus S$ denote the graph $G$ with set of vertices $S$ removed. Then with probability at least $1 - n^{-1+o(1)}$,
\[
\sum_{j \in \Gamma_1^G(i)} 1\{|\Gamma_\ell^G(j)\cap \delta_0| \geq \eta\} \geq d_i - |S|, \quad \forall S \text{ s.t. } i \in S, |S| \leq 3.
\]

Proof. We first show that with high probability for all $S$ and $j$, if $|\Gamma_1^G(j)|$ is large, then $|\Gamma_\ell^G(j)|$ grows in depth $\ell$ by at least a multiplicative factor of $np/2$. Define event
\[
A = \left\{ \left| \Gamma_\ell^G(j) \right| \geq \frac{2(np/2)^\ell}{\log(np)}, \forall S, \forall j \text{ s.t. } |S| \leq 3 \text{ and } \frac{np}{\log(np)} \leq |\Gamma_1^G(j)| \leq 4np \right\}.
\]
For any fixed set $S, G \setminus S \sim G(n - |S|, p)$. Under the choice of $\ell$ in (17), $(4np)^\ell = o(n)$. Applying Corollary 1 together with union bounds over all possible choices of $S$ and $j$, we get that
\[
\mathbb{P}\{A\} \geq 1 - n \left( \sum_{k=0}^{3} \binom{n}{k} \right) \exp \{-\Omega\left( (np)^2 / \log(np) \right) \} \geq 1 - n^{-o(1)}.
\]

Then we show with high probability for all $i$ and $S$, vertex $i$ has at most $|S|$ different neighbors $j$ whose $|\Gamma_1^G(j)|$ is small. Define event
\[
B = \left\{ \sum_{j \in \Gamma_1^G(i)} 1\{|\Gamma_\ell^G(j)| \leq \frac{np}{\log(np)}\} \leq |S|, \forall S, \forall i \text{ s.t. } i \in S, |S| \leq 3 \right\}
\]
and event
\[
\tilde{B} = \left\{ \sum_{j \in \Gamma_1^G(i)} 1\{|\Gamma_\ell^G(j)| \leq \frac{np}{\log(np)} + 3\} \leq 1, \forall i \right\}.
\]

Let $\mathcal{E}$ denote the event that the maximum degree in $G$ is at most $4np$ as defined in Lemma 10. In view of Lemma 16, we have that $\mathbb{P}\{\tilde{B} \cap \mathcal{E}\} \geq 1 - n^{-1+o(1)}$. Now we claim that $\tilde{B} \subset B$. To see this, among all $j \in \Gamma_1^G(i)$, there are at most $|S| - 1$ different $j \in S$. Moreover, for $j \notin S$, $|\Gamma_1^G(j)| \leq |\Gamma_1^G(j)| + |S|$ and thus
\[
1\{|\Gamma_\ell^G(j)| \leq \frac{np}{\log(np)}\} \leq 1\{|\Gamma_1^G(j)| \leq \frac{np}{\log(np)} + 3\}.
\]
As a consequence, $\tilde{B} \subset B$ and it follows that $\mathbb{P}\{B \cap \mathcal{E}\} \geq 1 - n^{-1+o(1)}$.

Recall that $\ell = \left(\frac{(1-\epsilon)\log n}{\log(np)}\right)$ and $\eta = 4^{2\ell + 2p^{2\ell} n^{2\ell - 1} \alpha}$. Then for sufficiently large $n$,
\[
\frac{2(np/2)^\ell}{\log(np)} \geq \frac{4\eta}{\alpha}.
\]
Hence, on event $A \cap B \cap \mathcal{E}$,
\[
\sum_{j \in \Gamma_1^G(i)} 1\{|\Gamma_\ell^G(j)| \geq 4\eta/\alpha\} \geq d_i - |S|, \quad \forall i, S \text{ s.t. } i \in S, |S| \leq 3.
\]
Finally, we show that with high probability for all $S$ and $j$, if $|\Gamma_\epsilon^{G \setminus S}(j)|$ is large, then $\Gamma_\epsilon^{G \setminus S}(j)$ contains many seeded vertices. Define event
\[
\mathcal{X} = \left\{ \left| \Gamma_\epsilon^{G \setminus S}(j) \cap I_0 \right| \geq \eta, \forall S, \forall j \text{ s.t. } |S| \leq 3 \text{ and } \left| \Gamma_\epsilon^{G \setminus S}(j) \right| \geq 4\eta/\alpha \right\}.
\]

Hence, on event $\mathcal{A} \cap B \cap \mathcal{E} \cap \mathcal{X}$, we have
\[
\sum_{j \in \Gamma_\epsilon^{G}(i)} 1 \left\{ \left| \Gamma_\epsilon^{G \setminus S}(j) \cap I_0 \right| \geq \eta \right\} \geq d_i - |S|, \quad \forall i, S \text{ s.t. } i \in S, |S| \leq 3.
\]

It remains to show $\mathbb{P} \{ \mathcal{A} \cap B \cap \mathcal{E} \cap \mathcal{X} \} \geq 1 - n^{-1+o(1)}$, which further reduces to proving $\mathbb{P} \{ \mathcal{X}^c \} \leq n^{-1+o(1)}$ by the union bound. Note that
\[
\mathcal{X}^c = \bigcup_S \bigcup_j \left\{ \left| \Gamma_\epsilon^{G \setminus S}(j) \right| \geq 4\eta/\alpha, \left| \Gamma_\epsilon^{G \setminus S}(j) \cap I_0 \right| < \eta \right\}.
\]
Thus by the union bound,
\[
\mathbb{P} \{ \mathcal{X}^c \} \leq \sum_S \sum_j \mathbb{P} \left\{ \left| \Gamma_\epsilon^{G \setminus S}(j) \right| \geq 4\eta/\alpha, \left| \Gamma_\epsilon^{G \setminus S}(j) \cap I_0 \right| < \eta \right\}
\leq \sum_S \sum_j \mathbb{P} \{ \text{Bin} \left( \left\lfloor \frac{4\eta}{\alpha} \right\rfloor, \alpha \right) \leq \eta \}
\leq n^4 e^{-\eta} \leq n^{-2},
\]
where the second inequality holds because each vertex is included in $I_0$ with probability $\alpha$ independently of everything else; the last inequality follows from the Binomial tail bound (F.1) and the assumption that $\eta \geq 6 \log n$.

The second lemma is useful to conclude that the test statistic $Z_{u,v}$ given in (15) 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 $\epsilon$ such that
\[
\sum_{i \in \Gamma_\epsilon^{G}(u)} \sum_{j \in \Gamma_\epsilon^{G}(v)} 1 \left\{ \left| \Lambda_\epsilon^{G \setminus (u)}(i) \cap \Lambda_\epsilon^{G \setminus (v)}(j) \cap I_0 \right| \geq \eta \right\} \leq C.
\]

**Proof.** For two vertices $i, j$, define
\[
c_{ij} = 1 \left\{ \left| \Lambda_\epsilon^{G \setminus (i)}(j) \cap \Lambda_\epsilon^{G \setminus (j)}(i) \right| \geq \eta/(4\alpha) \right\}
\]
and event
\[
C = \left\{ \max_i \sum_j c_{ij} \leq 2n^{4\epsilon} \right\} \cap \left\{ \max_j \sum_i c_{ij} \leq 2n^{4\epsilon} \right\}.
\]
In view of Lemma 14 and $\eta = 4^{2\epsilon+2} p^{2\epsilon} n^{-\epsilon-1} \alpha$, $\mathbb{P} \{ C \} \geq 1 - 2/n$. 

Define

\[ a_{ij} = 1_{\{ |N_i^c(0) \cap N_j^c(0) \cap I_0| \geq \eta \}} \]

and event

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

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 \( c_{ij} = 0 \), it holds that \( a_{ij} = 0 \); thus \( a_{ij} \leq c_{ij} \) for all \( i, j \). Hence, \( C \cap \mathcal{Y}^c \subset \mathcal{A} \) and thus

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

Note that

\[ \mathbb{P} \{ c_{ij} = 0, a_{ij} = 1 \} \leq \mathbb{P} \{ a_{ij} = 1 \mid c_{ij} = 0 \} \leq \mathbb{P} \{ \text{Bin}(\lfloor \eta/4\alpha \rfloor, \alpha) \geq \eta \} \leq e^{-2\eta}. \]

By the union bound, we have

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

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

Fix a pair of vertices \( u \neq v \) in the sequel, and let

\[ b_{ij} = 1_{\{ |N_i^c(u) \cap N_j^c(v) \cap I_0| \geq \eta \}} \]

and

\[ B_{u,v} = \left\{ \max_i \sum_j b_{ij} \leq 2n^{4e} \right\} \cap \left\{ \max_j \sum_i b_{ij} \leq 2n^{4e} \right\} . \]

Then by construction, \( b_{ij} \leq a_{ij} \) and thus \( \mathcal{A} \subset B_{u,v} \).

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

\[ R_{u,v} = \sum_{i,j \in [n] \setminus \{u,v\}} b_{ij}X_iX_{n+j}, \]

which is a degree-2 polynomial of \( X_i \)'s. Note that \( \{ b_{ij} \mid i, j \in [n] \setminus \{u,v\} \} \) only depends on \( G \setminus \{u,v\} \) 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 \( B_{u,v} \) holds. Let

\[ \mu_0 = \mathbb{E} \left[ R_{u,v} \mid b \right] = p^2 \sum_{ij} b_{ij} \leq 2p^2 n^{1+4e} \leq 2n^{-1+6e} \]
and
\[ \mu_1 = \max \left\{ \max_i E \left[ \sum_j b_{ij} X_j \mid b \right], \max_j E \left[ \sum_i b_{ij} X_i \mid b \right] \right\} \leq 2pn^4 \leq 2n^{-1+5\epsilon}, \]

where we used the fact that \( np \leq n^\epsilon \).

By a concentration inequality for multivariate polynomials [38, Corollary 4.9], there exists a constant \( C > 0 \) depending only on \( \epsilon \) such that
\[ P\left\{ R_{u,v} \geq C \mid b \right\} \leq n^{-3}. \]

Define event \( E_{u,v} = \{ R_{u,v} \geq C \} \) and \( \mathcal{R} = \bigcup_{u,v} R_{u,v} \). It follows that
\[ P\left\{ R_{u,v} \cap B_{u,v} \right\} = \sum_{b \in B_{u,v}} P\left\{ R_{u,v} \mid b = \tilde{b} \right\} \leq n^{-3} \sum_{b \in B_{u,v}} P\left\{ b = \tilde{b} \right\} \leq n^{-3}. \]

Since \( \mathcal{A} \subset B_{u,v} \), it further follows that \( P\{ R_{u,v} \cap \mathcal{A} \} \leq n^{-3} \). By a union bound over \( u,v \), we have
\[ P\{ \mathcal{R} \cap \mathcal{A} \} \leq P\{ \mathcal{R} \} + P\{ \mathcal{A}^c \} \leq 4/n. \]

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

**Proof of Theorem 3.** Recall that \( \tau = \mu_1 \) is given in (12) 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 that \( \ell \) and \( \eta \) are given in (17) and (18), respectively. Recall the definition of \( Z \) given in (15).

By the theorem assumptions, \( \eta \geq n^{2\epsilon} \geq 6 \log n \) for all sufficiently large \( n \). Applying Lemma 8 with \( G = G_1^* \wedge G_2 \), we get that with high probability, for all high-degree vertices \( u \),
\[ Z_{u,u} \geq \tau - 3 = \frac{npn^2}{\log(npn^2)} - 3. \]

Moreover, by definition,
\[ w_{ij}^{u,v} \leq \left| \left\{ k \in I_0 : \pi_0(k) \in N_{G_1^* \setminus \{u,v\}}(i), k \in N_{G_2^* \setminus \{u,v\}}(j) \right\} \right| \]
\[ \leq \left| N_{G_1^* \wedge G_2^* \setminus \{u,v\}}(i) \cap N_{G_1^* \wedge G_2^* \setminus \{u,v\}}(j) \cap I_0 \right|. \]

Applying Lemma 9 with \( G = G_1^* \vee G_2 \), we get that with high probability,
\[ Z_{u,v} \leq C, \quad \forall u \neq v \]
for a constant \( C > 0 \) only depending on \( \epsilon \). Since for sufficiently large \( n \), \( \tau \geq C + 4 \), 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.

**ACKNOWLEDGMENTS**

The authors would like to thank the anonymous reviewers for constructive comments. 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 is supported by the NSF Grants IIS-1932630, CCF-1850743, and CCF-1856424.

REFERENCES

1. N. Alon and J.H. Spencer, *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization, 3rd ed., Wiley, Hoboken, NJ, 2008.

2. L. Babai, *Graph isomorphism in quasipolynomial time*. Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing, STOC ’16, ACM, New York, 2016, pp. 684–697.

3. B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng, *Efficient algorithms for the graph matching problem on correlated random graphs*, arXiv preprint arXiv:1805.02349, 2018.

4. R.E. Burkard, E. Cela, P.M. Pardalos, and L.S. Pitsoulis, *The quadratic assignment problem*. Handbook of combinatorial optimization, Springer, Berlin, pp. 1713–1809, 1998.

5. L. Babai, P. Erdős, and S.M. Selkow, *Random graph isomorphism*, SIAM J. Comput. 9 (1980), 628–635.

6. L. Babai, D.Y. Grigoryev, and D.M. Mount, *Isomorphism of graphs with bounded eigenvalue multiplicity*. Proceedings of the fourteenth annual ACM symposium on Theory of computing, ACM, New York, pp. 310–324, 1982.

7. B. Bollobás, *Distinguishing vertices of random graphs*, North-Holland Math. Stud. 62 (1982), 33–49.

8. B. Bollobás, *Random graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics, Cambridge, MA, 2001.

9. D. Conte, P. Foggia, C. Sansone, and M. Vento, *Thirty years of graph matching in pattern recognition*, Int. J. Pattern Recogn. Artif. Intell. 18 (2004), 265–298.

10. D. Cullina and N. Kiyavash, *Improved achievability and converse bounds for Erdős-Rényi graph matching*. Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, ACM, New York, pp. 63–72, 2016.

11. D. Cullina and N. Kiyavash, *Exact alignment recovery for correlated Erdős-Rényi graphs*, arXiv preprint arXiv:1711.06783, 2017.

12. D. Cullina, N. Kiyavash, P. Mittal, and H. Vincent Poor. *Partial recovery of Erdős-Rényi graph alignment via k-core alignment*, arXiv preprint arXiv:1809.03553, 2018.

13. T. Czajka and G. Pandurangan, *Improved random graph isomorphism*, J. Discr. Algorithms. 6 (2008), 85–92.

14. O.E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser, *Analysis of a canonical labeling algorithm for the alignment of correlated Erdős-Rényi graphs*, arXiv preprint arXiv:1804.09758, 2018.

15. J. Ding, Z. Ma, Y. Wu, and J. Xu, *Efficient random graph matching via degree profiles*, arXiv preprint arXiv:1811.07821, 2018.

16. D.E. Fishkind, S. Adali, and C.E. Priebe, *Seeded graph matching*, arXiv preprint arXiv:1209.0367, 2018.

17. L.R. Ford and D.R. Fulkerson, *Maximal flow through a network*, Can. J. Math. 8 (1956), 399–404.

18. A. Frieze and M. Karoński, *Introduction to random graphs*, Cambridge University Press, Cambridge, MA, 2015.

19. S. Feizi, G. Quon, M. Recamonde-Mendoza, M. Médard, M. Kellis, and A. Jadbabaie, *Spectral alignment of networks*, arXiv preprint arXiv:1602.04181, 2016.

20. A.D. Haghighi, A.Y. Ng, and C.D. Manning, *Robust textual inference via graph matching*. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 387–394, 2005.

21. S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, John Wiley & Sons, Inc., New York, 2011.

22. E. Kazemi, S.H. Hassan, and M. Grossglauser, *Growing a graph matching from a handful of seeds*, Proc. VLDB Endow. 8 (2015), 1010–1021.

23. N. Korula and S. Lattanzi, *An efficient reconciliation algorithm for social networks*, Proc. VLDB Endow. 7 (2014), 377–388.

24. V. Lyzinski, D.E. Fishkind, and C.E. Priebe, *Seeded graph matching for correlated Erdős-Rényi graphs*, J. Mach. Learn. Res. 15 (2013).

25. R.J. Lipton, *The beacon set approach to graph isomorphism*, Technical report, Yale University (1978).

26. L. Livi and A. Rizzi, *The graph matching problem*, Pattern Anal. Appl. 16 (2013), 253–283.

27. J. Lubars and R. Srikant, *Correcting the output of approximate graph matching algorithms*. IEEE INFOCOM 2018-IEEE Conference on Computer Communications, IEEE, New York, pp. 1745–1753, 2018.
APPENDIX A | PROOF OF THEOREM 2

Proof of Theorem 2. Suppose \( np^s - \log n = c < +\infty \). Since \( G^*_1 \wedge G_2 \sim G(n, ps^2) \), classical random graph theory shows that the distribution of the number of isolated vertices in \( G^*_1 \wedge G_2 \) converges to \( \text{Pois}(e^{-c}) \), see, for example, [8, Theorem 3.1]. Let \( F_1 \) denote the event that there are at least two isolated vertices in \( G^*_1 \wedge G_2 \). Then \( \mathbb{P} \{ F_1 \} = \Omega(1) \).

Let \( F_2 \) 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 \( \alpha \) independently across different vertices and from the graphs \( G_1 \) and \( G_2 \), it follows that \( \mathbb{P} \{ F_2 \} \geq \mathbb{P} \{ F_1 \} (1 - \alpha)^2 = \Omega \left( (1 - \alpha)^2 \right) \).

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

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

\[
\hat{\pi}_{\text{ML}} \in \arg \min_{\pi \in S} \min_{\Pi} \| G_1 - \Pi G_2 \Pi^\top \|_F^2,
\]
where $\Pi$ is the permutation matrix corresponding to permutation $\pi$; or equivalently,

$$\hat{\pi}_{\text{ML}} \in \arg\max_{\pi \in \mathcal{S}} \left\langle G_1, \Pi G_2 \Pi^T \right\rangle.$$ 

Let $I$ denote the union of the initial seed set and the set of all nonisolated vertices in $G^*_1 \land G_2$. Then $I^c$ is the set of isolated vertices that are unseeded in $G^*_1 \land G_2$. Let $\tilde{S}$ denote the set of all possible permutations $\pi$ such that $\pi(i) = \pi^*(i)$ for $i \in I$. Then $\pi^* \in \tilde{S} \subset S$. Note that for any $\pi \in \tilde{S}$, we have

$$\left\langle G_1, \Pi G_2 \Pi^T \right\rangle \geq \sum_{(i,j) \notin I \times I} G_1(\pi(i), \pi(j)) G_2(i, j)$$

$$= \sum_{(i,j) \in I \times I} G_1(\pi^*(i), \pi^*(j)) G_2(i, j)$$

$$= \sum_{(i,j) \in I \times I} G_1(\pi^*(i), \pi^*(j)) G_2(i, j),$$

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 least $|I^c| \geq |I^c| - 1$ different permutations in $\tilde{S}$ whose likelihood is at least as large as the ground truth $\pi^*$, and hence the MLE is correct with probability at most $1/(|I^c| - 1)$. Note that on event $F_2$, $|I^c| \geq 2$; hence, MLE is correct with probability at most $1/2$. In conclusion, MLE is correct with probability at most $(1/2) \mathbb{P} \{ F_2 \} = \Omega((1 - \alpha)^2)$.

### 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

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

Note that

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

$$\leq (\mathbb{P} \{\text{Bin}(n - 2, p) \leq \tau\})^2$$

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

$$\mathbb{P} \{\text{Bin}(n - 2, p) \leq \tau\} \leq \exp \left( -\frac{\tau}{np} \right) \leq \exp \left( -\frac{\sqrt{\tau}}{n} \right) = \exp \left( -(1 - o(1))np \right).$$
Combining the last two displayed equations yields that
\[
P \{ \mathcal{E}_{ij} \} = P \{ G(i,j) = 1 \} \ P \{ d(i) \leq \tau, d(j) \leq \tau | G(i,j) = 1 \} \leq p \exp(-2(1 - o(1))np)\]

By the union bound,
\[
P \{ \bigcup_{i \neq j} \mathcal{E}_{ij} \} \leq n^2 P \{ \mathcal{E}_{ij} \} \leq n^2 p \exp(-2(1 - o(1))np) \leq n^{-1 + o(1)},\]

where the last inequality holds due to \( np - \log n \to +\infty \).

\section*{Appendix C \ | Proof of Lemma 2}

\textit{Proof.} Recall that we have assumed \( \pi^* = id \) and thus \( G_1^* = G_1 \). Let \( d_i \) denote the degree of vertex \( i \) in \( G_1 \land G_2 \) and \( A \) denote the adjacency matrix of \( G_1 \lor G_2 \). For every pair of three distinct vertices \( i, j, k \), define
\[
\mathcal{F}_{ijk} = \{ A_{ik} = 1, A_{jk} = 1 \} \cap \{ d_i \leq \tau \} \cap \{ d_j \leq \tau \}.
\]

It suffices to show that \( P \{ \bigcup_{i,j,k} \mathcal{F}_{ijk} \} \leq n^{-1 + o(1)} \). Since \( G_1 \lor G_2 \sim \mathcal{G}(n, ps(2 - s)) \), it follows that
\[
P \{ A_{ik} = 1, A_{jk} = 1 \} = P \{ A_{ik} = 1 \} P \{ A_{jk} = 1 \} = (ps(2 - s))^2 \leq p^2.
\]

Moreover, since \( G_1 \land G_2 \sim \mathcal{G}(n, ps^2) \), it follows that
\[
P \{ \{ d_i \leq \tau \} \cap \{ d_j \leq \tau \} | A_{ik} = 1, A_{jk} = 1 \} \leq \left( P \{ \text{Binom}(n-3, ps^2) \leq \tau \} \right)^2.
\]

In view of Binomial tail bound (F.1) and \( \tau = o(nps^2) \), we have that
\[
P \{ \text{Binom}(n-3, ps^2) \leq \tau \} \leq \exp \left(-\frac{(n-3)ps}{\sqrt{n-3}ps^2} \right)^2
= \exp (-nps^2 (1 - o(1)))
\]

It follows that
\[
P \{ \mathcal{F}_{ijk} \} \leq p^2 \exp(-2nps^2 (1 - o(1)))
\]

By the union bound, we have that
\[
P \{ \bigcup_{i,j,k} \mathcal{F}_{ijk} \} \leq n^3 p^2 \exp(-2nps^2 (1 - o(1))) = n^{-1 + o(1)}.
\]

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

\section*{Appendix D \ | Neighborhood Exploration in \( \mathcal{G}(n, p) \)}

Throughout this section, we assume graph \( G \sim \mathcal{G}(n, p) \) with \( np \geq \log n \). We first claim that the max degree in \( G \) is at most \( 4np \) with probability at least \( 1 - 1/n \).
Lemma 10. Assume graph $G \sim G(n, p)$ with $np \geq \log n$. Let

$$\mathcal{E} = \left\{ \max_{v \in V(G)} d_v \leq 4np \right\}. \tag{D.1}$$

Then

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

Proof. By the Binomial tail bound (F.2),

$$\mathbb{P}\{d_v \geq 4np\} = \mathbb{P}\{\text{Binom}(n - 1, p) \geq 4np\} \leq \exp(-2np).$$

The lemma follows by a union bound over $v \in V(G)$ and the assumption that $np \geq \log n$. \hfill \Box

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 $np \geq n^\epsilon$.

APPENDIX D.1 Dense regime

The following lemma is adapted from [8, 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 $u$,

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

(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 $8n^{2d-3}p^{2d-2}$. 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 $w \not\in N_{d-2}(u) \cup N_{d-2}(v)$, 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 $2n[2p(np)^{d-2}]^2 = 8p^{2d-2}n^{2d-3}$.
APPENDIX D.2 | Sparse regime

In contrast, in the sparse regime where

\[ np - \log n \to +\infty. \]

there exist vertices with small degrees, that is, \(|\Gamma_1(u)|\) 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 \(|\Gamma_k(u)| \approx (np)|\Gamma_{k-1}(u)|\) for all \(2 \leq k \leq d\) for some \(d\) with high probability.

**Lemma 12.** Suppose

\[ np \geq \log n \quad \text{and} \quad p(4np)^{d-1} = o(1). \quad (D.2) \]

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

\[ Q_k = \left\{ |\Gamma_k(u)| \in l_k = \left[ \tau \left( \frac{np}{2} \right)^{k-1}, (4np)^k \right] \right\} \]

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

\[ \mathbb{P}\{Q_k \mid Q_1, \ldots, Q_{k-1}\} \geq 1 - \exp \left( -\Omega \left( \frac{np}{2} \right)^{k-1} \right). \]

It readily follows that

\[ \mathbb{P}\{Q_d \cap Q_{d-1} \cap \cdots \cap Q_2 \mid Q_1\} \geq 1 - \exp (\Omega(np)). \]

**Proof.** Fix \(2 \leq k \leq d\). Conditional on \(\Gamma_{k-1}(u)\) and \(N_{k-1}(u)\), a given vertex \(v \notin N_{k-1}(u)\) connects to some vertices in \(\Gamma_{k-1}(u)\) with probability

\[ p_k = 1 - (1 - p)^{|\Gamma_{k-1}(u)|}. \]

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

\[ |\Gamma_k(u)| \sim \text{Bin}(n - |N_{k-1}(u)|, p_k). \]

Note that conditional on \(Q_1, \ldots, Q_{k-1}\),

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

where the last equality holds due to the assumption (D.2) and \(k \leq d\). Moreover, in view of the assumption (D.2), conditional on \(Q_1, \ldots, Q_{k-1}\),

\[ (1 - o(1)) \frac{np}{2} \left( \frac{np}{2} \right)^{k-2} \leq p_k \leq p(4np)^{k-1}. \]
Hence, for $2 \leq k \leq d$,

$$\mathbb{P}\{|\Gamma_k(u)| \notin I_k \mid Q_1, \ldots, Q_k\} \leq \mathbb{P}\left\{ \text{Bin} \left( n - o(n), \left(1 - o(1)\right) p \tau \left( \frac{np}{2} \right)^{k-1} \right) \leq \tau \left( \frac{np}{2} \right)^{k-1} \right\} + \mathbb{P}\left\{ \text{Bin} \left(n, p(4np)^{k-1}\right) \geq (4np)^k \right\} \leq \exp \left(-\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right) + \exp \left(-4^{k-1}(np)^k\right) \leq \exp \left(-\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right).$$

Finally, we note that

$$\mathbb{P}\{Q_d \cap Q_{d-1} \cap \cdots \cap Q_2 \mid Q_1\} = \mathbb{P}\{Q_2 \mid Q_1\} \mathbb{P}\{Q_3 \mid Q_1, Q_2\} \cdots \mathbb{P}\{Q_d \mid Q_1, \ldots, Q_{d-1}\} \geq \prod_{k=2}^{d-1} \left( 1 - \exp \left(-\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right) \right) \geq 1 - \sum_{k=2}^{d-1} \exp \left(-\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right) \geq 1 - \exp(-\Omega(\tau np)).$$

With Lemma 12, we have the following immediate corollary.

**Corollary 1.** Suppose (D.2) holds for some integer $d \geq 1$. Define event

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

Then

$$\mathbb{P}\{Q\} \geq 1 - n \exp(-\Omega(\tau np)).$$

**Proof.** Note that

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

Hence, it follows from the union bound that

$$\mathbb{P}\{Q^c\} \leq \sum_u \mathbb{P}\{\{\tau \leq |\Gamma_1(u)| \leq 4np\} \cap \{|\Gamma_k(u)| \notin I_k, \ \forall 1 \leq k \leq d\} \} \leq \sum_u \mathbb{P}\{|\Gamma_k(u)| \notin I_k, \ \forall 1 \leq k \leq d \mid \tau \leq |\Gamma_1(u)| \leq 4np\} \leq n \exp(-\Omega(\tau np)), $$

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
\[
\Gamma^*_k(u, v) = \{ w \in \Gamma_k(u) \cap \Gamma_k(v) : \Gamma_1(w) \cap (\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)) \neq \emptyset, \Gamma_1(w) \cap (\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)) \neq \emptyset \}
\]
and we abbreviate $\Gamma^*_k(u, v)$ as $\Gamma^*_k(u, v)$ for simplicity. By definition, for any $d \geq 1$,
\[
\Gamma_d(u) \cap \Gamma_d(v) \subset \bigcup_{k=1}^d \Gamma_{d-k}(\Gamma^*_k(u, v))
\]
and
\[
N_d(u) \cap N_d(v) \subset \bigcup_{\ell=-d}^d \bigcup_{k=0}^d N_{d-k-\max(\ell, 0)}(\Gamma^*_k(u, v)).
\]
The following lemma gives an upper bound to $|\Gamma^*_k(u, v)|$ in high probability.

**Lemma 13.** For two distinct vertices $u, v$, define for $k, \ell \geq 1$,
\[
\Delta_{k, \ell} = \{ |\Gamma_{k-1}(u)| \leq (4np)^{k-1}, |\Gamma_{\ell-1}(v)| \leq (4np)^{\ell-1} \}.
\]

Then
\[
\mathbb{P} \left\{ \left| \Gamma^*_k(u, v) \right| \geq \gamma_{k+\ell} \mid \Delta_{k, \ell} \right\} \leq n^{-8}, \tag{D.3}
\]
where
\[
\gamma_k = \begin{cases} 24 \log n & \text{if } np^2 (4np)^{k-2} \leq 4 \log n \\ 4np^2 (4np)^{k-2} & \text{o.w.} \end{cases} \tag{D.4}
\]

**Proof.** Conditional on $(N_{k-1}(u), \Gamma_{k-1}(u))$ and $(N_{\ell-1}(v), \Gamma_{\ell-1}(v))$, a given vertex $w \not\in N_{k-1}(u) \cup N_{\ell-1}(v)$ connects to some vertex in $\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)$ with probability
\[
1 - (1 - p)^{|\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)|} \leq p|\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)| \leq p|\Gamma_{k-1}(u)|.
\]

Similarly, $w$ connects to some vertex in $\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)$ with probability
\[
1 - (1 - p)^{|\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)|} \leq p|\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)| \leq p|\Gamma_{\ell-1}(v)|.
\]

Since $\Gamma_{k-1}(u) \setminus \Gamma_{\ell-1}(v)$ is disjoint from $\Gamma_{\ell-1}(v) \setminus \Gamma_{k-1}(u)$, it follows that $w \not\in \Gamma_{u,v}$ with probability at most $p^2 |\Gamma_{k-1}(u)||\Gamma_{\ell-1}(v)|$. Moreover, there are at most $n$ vertices $w \not\in N_{k-1}(u) \cup N_{\ell-1}(v)$. Hence,
\[
\mathbb{P} \left\{ \left| \Gamma^*_k(u, v) \right| \geq \gamma_{k+\ell} \mid \Delta_{k, \ell} \right\} \leq \mathbb{P} \left\{ \text{Bin} \left( n, p^2 (4np)^{k+\ell-2} \right) \geq \gamma_{k+\ell} \right\}.
\]

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 (F.3) that
\[
\mathbb{P} \left\{ \text{Bin} \left( n, p^2 (4np)^{k+\ell-2} \right) \geq \gamma_{k+\ell} \right\} \leq 2^{-\gamma_{k+\ell}} = 2^{-24 \log n} \leq n^{-8}.
\]
If \( np^2(4np)^{k+\varepsilon-2} \geq 4 \log n \), then by the choice of \( \gamma_{k+\varepsilon} = 4np^2(4np)^{k+\varepsilon-2} \), it follows from (F.2) that
\[
\Pr \left\{ \text{Bin} \left(n, p^2(4np)^{k+\varepsilon-2} \right) \geq \gamma_{k+\varepsilon} \right\} \leq \exp \left(-2np^2(4np)^{k+\varepsilon-2}\right) \leq n^{-8}.
\]

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

**Lemma 14.** Suppose that for a small constant \( \varepsilon > 0 \) and an integer \( 1 \leq d \leq n \),
\[
\log n \leq np \leq n^\varepsilon \quad \text{and} \quad (4np)^d \geq n^{1-\varepsilon}.
\]

For each vertex \( u \), define event
\[
\mathcal{R}_u = \left\{ \sum_{v} 1_{\{|N_d(u) \cap N_d(v)| > 4^{d+1}p^2n^{d-1}\}} \leq 2n^{4\varepsilon} \right\}
\]
and \( \mathcal{R} = \cap_u \mathcal{R}_u \). Then
\[
\Pr \{ \mathcal{R} \} \geq 1 - 2n^{-1}.
\]

**Proof.** Define an event
\[
\mathcal{A} = \cap_{u \neq v} \cap_{1 \leq k \leq d} \cap_{1 \leq \ell \leq d} \left\{ |\Gamma^{+}_{k,\ell}(u,v)| \leq \gamma_{k+\varepsilon} \right\}.
\]
Recall \( \mathcal{E} \) defined in (D.1). Note that
\[
(\mathcal{A} \cap \mathcal{E})^c = (\mathcal{A}^c \cap \mathcal{E}) \cup \mathcal{E}^c.
\]
Therefore,
\[
\Pr \{ (\mathcal{A} \cap \mathcal{E})^c \} \leq \Pr \{ \mathcal{A}^c \cap \mathcal{E} \} + \Pr \{ \mathcal{E}^c \}.
\]
In view of Lemma 10, \( \Pr \{ \mathcal{E}^c \} \leq 1/n \). Moreover,
\[
\Pr \{ \mathcal{A}^c \cap \mathcal{E} \} \leq \sum_{u \neq v} \sum_{1 \leq k \leq d} \sum_{1 \leq \ell \leq d} \Pr \left\{ \left| \Gamma^{+}_{k,\ell}(u,v) \right| \geq \gamma_{k+\varepsilon} \right\} \cap \mathcal{E} \}
\]
\[
\leq (a) \sum_{u \neq v} \sum_{1 \leq k \leq d} \sum_{1 \leq \ell \leq d} \Pr \left\{ \left| \Gamma^{+}_{k,\ell}(u,v) \right| \geq \gamma_{k+\varepsilon} \right\} \cap \Delta_{k,\ell} \}
\]
\[
\leq \sum_{u \neq v} \sum_{1 \leq k \leq d} \sum_{1 \leq \ell \leq d} \Pr \left\{ \left| \Gamma^{+}_{k,\ell}(u,v) \right| \geq \gamma_{k+\varepsilon} \right\} \cap \Delta_{k,\ell} \}
\leq n^{-d},
\]
where \( (a) \) follows from \( \mathcal{E} \subset \Delta_{k,\ell} \) and the last inequality holds in view of Lemma 13 and \( d \leq n \). Therefore, \( \Pr \{ (\mathcal{A} \cap \mathcal{E})^c \} \leq 2/n \).

To prove the lemma, it remains 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
\[
N_d(u) \cap N_d(v) \subset \cup_{\ell = d}^{d-1} \cup_{k=0}^{d-\max\{\ell,0\}} \Gamma^{+}_{k,\ell}(u,v).
\]
It follows that

\[ |N_d(u) \cap N_d(v)| \leq \sum_{\ell = -d}^{d} \sum_{k=0}^{d} \mathbf{1}_{\{0 \leq k + \ell \leq d\}} \left| \Gamma^*_{k+\ell}(u, v) \right| (4np)^{d-k-\max\{\ell, 0\}}. \]

Set

\[ k_0 = \left\lfloor \frac{2\varepsilon \log n}{\log(4np)} \right\rfloor. \]

Then

\[ |N_{2k_0}(u)| \leq \sum_{k=0}^{2k_0} (4np)^k \leq \frac{(4np)^{2k_0+1} - 1}{4np - 1} \leq 2(4np)^{2k_0} \leq 2n^{4\varepsilon}, \]

where the second-to-the-last inequality holds due to \( 2np \geq 1 \). Note that for all \( v \not\in N_{2k_0}(u) \),

\[ |\Gamma^*_{k,\ell}(u, v)| = 0, \quad \forall \ 0 \leq k + \ell \leq 2k_0. \]

and hence,

\[ |N_d(u) \cap N_d(v)| \leq \sum_{\ell = -d}^{d} \sum_{k=0}^{d} \mathbf{1}_{\{0 \leq k + \ell \leq d\}} \mathbf{1}_{\{2k+\ell \geq 2k_0+1\}} \left( 24 \log n + 4np^2(4np)^{2k+\ell-2} \right) (4np)^{d-k-\max\{\ell, 0\}} \]

\[ \leq 192 \log n (4np)^{d-k_0-1/2} + 32np^2(4np)^{2d-2} \]

\[ \leq 64np^2(4np)^{2d-2} = 4^{d+1}p^{2d}n^{2d-1}, \]

where the last inequality holds because by assumption (D.5),

\[ (4np)^{d+k_0+1/2} \geq \frac{1}{2} n^{1+\varepsilon/2} \geq 96n \log n \]

for \( n \) sufficiently large. Hence, for every \( u \),

\[ \sum_{v} \mathbf{1}_{\{|N_d(u) \cap N_d(v)| > 4^{d+1}p^{2d}n^{2d-1}\}} \leq |N_{2k_0}(u)| \leq 2n^{4\varepsilon}. \]

As a consequence, \( \mathcal{A} \cap \mathcal{E} \subset \mathcal{R} \).

**APPENDIX D.3 | Graph branching in sparse regime**

In this subsection, we describe a branching process to explore the vertices in \( N_k(u) \). See, for example, [1, 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 popped 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 popped 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 subree $T_k(v)$. 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 children $v$ of root $u$ such that $|\Pi_1(v)| \geq \tau$, $|\Pi_k(v)|$ grows at least as $\tau (np/2)^{k-1}$ for $k$ from 1 up to some depth $d$.

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

$$F_1 = \{|\Pi_1(u)| \leq 4np\} \cap \{|\Pi_1(v)| \leq 4np, \forall v \in \Pi_1(u)\},$$

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

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

Suppose

$$np \geq \log n \quad \text{and} \quad (4np)^{d+1} = o(n). \quad (D.7)$$

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

$$\mathbb{P}\{F_k | F_1, \ldots, F_{k-1}\} \geq 1 - 8np \exp\left(-\Omega\left(\tau \left(\frac{np}{2}\right)^{k-1}\right)\right).$$

It further follows that

$$\mathbb{P}\{F_d \cap F_{d-1} \cap \cdots \cap F_2 | F_1\} \geq 1 - 8np \exp(-\Omega(np)).$$

Moreover, by letting

$$A_u = (F_d \cap F_{d-1} \cap \cdots \cap F_2) \cup F_1^c,$$

we have

$$\mathbb{P}\{A_u^c\} \leq 8np \exp(-\Omega(np)).$$

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

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

$$F_{k,i} = \{|\Pi_k(v_j)| \leq (4np)^k, \forall j \in [i]\} \cap \{|\Pi_k(v_j)| \geq \tau \left(\frac{np}{2}\right)^{k-1}, \forall j \in [i] \text{ s.t. } |\Pi_1(v_j)| \geq \tau\}.$$

Then $F_{k,1} \supset F_{k,2} \supset \cdots \supset F_{k,d_u} = F_k$. 


Conditional on $\Pi_{k-1}(v_i)$, a given neutral vertex $w$ connects to some vertex in $\Pi_{k-1}(v_i)$ with probability

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

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 Bin $(n, \text{p}[\Pi_{k-1}(v_i)])$ and hence

$$P \{ |\Pi_k(v_i)| \geq (4np)^k | F_1, \ldots, F_{k-1}, F_{k,i-1} \} \leq P \{ \text{Bin}(n, p(4np)^{k-1}) \geq (4np)^k \}$$

$$\leq \exp \left( -4(k-1)(np)^k \right),$$

(D.8)

where the last inequality follows from the Binomial tail bound (F.2).

On the other hand, in view of assumption (D.7), conditional on $F_1, \ldots, F_{k-1}, F_{k,i-1}$, there are at least

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

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

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

where the last inequality holds due to the assumptions that $\tau \leq np$ and $(4np)^d + 1 = o(n)$. Therefore, conditional on $F_1, \ldots, F_{k-1}, F_{k,i-1}, [\Pi_k(v_i)]$ is stochastically lower bounded by

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

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

$$P \left( |\Pi_k(v_i)| \leq \tau \left( \frac{np}{2} \right)^{k-1} | F_1, \ldots, F_{k-1}, F_{k,i-1} \right)$$

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

$$\leq \exp \left( -\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right).$$

(D.9)

Combining (D.8) and (D.9) yields that

$$P \{ F_{k,i} | F_1, \ldots, F_{k-1} \} = P \left( F_{k,i} \cap F_{k,i-1} | F_1, \ldots, F_k \right)$$

$$= P \{ F_{k,i-1} | F_1, \ldots, F_{k-1} \} P \{ F_{k,i} | F_1, \ldots, F_k, F_{k,i-1} \}$$

$$\geq P \{ F_{k,i-1} | F_1, \ldots, F_{k-1} \} \left( 1 - 2 \exp \left( -\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right) \right).$$

Therefore,

$$P \{ F_k | F_1, \ldots, F_{k-1} \} \geq 1 - 8np \exp \left( -\Omega \left( \tau \left( \frac{np}{2} \right)^{k-1} \right) \right).$$
Finally, we note that
\[
\mathbb{P}\{ \mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2 \mid \mathcal{F}_1 \} = \mathbb{P}\{ \mathcal{F}_2 \mid \mathcal{F}_1 \} \mathbb{P}\{ \mathcal{F}_3 \mid \mathcal{F}_1, \mathcal{F}_2 \} \cdots \mathbb{P}\{ \mathcal{F}_d \mid \mathcal{F}_1, \ldots, \mathcal{F}_{d-1} \} \\
\geq \prod_{k=2}^{d} \left( 1 - 8np \exp \left( -\Omega \left( \frac{np}{2} \right)^{k-1} \right) \right) \\
\geq 1 - 8np \sum_{k=2}^{d} \exp \left( -\Omega \left( \frac{np}{2} \right)^{k-1} \right) \\
\geq 1 - 8np \exp (-\Omega(np)).
\]

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,
\[
\mathbb{P}\{ \mathcal{A}_u^c \} = \mathbb{P}\{ \mathcal{F}_1 \} \mathbb{P}\{ (\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2)^c \mid \mathcal{F}_1 \} \\
\leq \mathbb{P}\{ (\mathcal{F}_d \cap \mathcal{F}_{d-1} \cap \cdots \cap \mathcal{F}_2)^c \mid \mathcal{F}_1 \} \\
\leq 8np \exp (-\Omega(np)),
\]
completing the proof.

The following lemma shows that with high probability, for all possible root vertex \( u \), it has at most one child \( v \) with \( |\Pi_1(v)| \leq \tau \) for \( \tau = o(np) \).

**Lemma 16.** Let \( A \) denote the adjacency matrix of \( G \sim G(n, p) \). Let \( u \) be the root vertex. For two distinct vertices \( v, w \), define
\[
\mathcal{B}_{u,v,w} = \{ A_{u,v} = 1, A_{u,w} = 1 \} \cap \{ |\Pi_1(v)| \leq \tau \} \cap \{ |\Pi_1(w)| \leq \tau \}.
\]

and \( B = \cup_{u,v,w} \mathcal{B}_{u,v,w} \). Assume
\[
np \geq \log n, \quad np = o(n^{1/2}), \quad \text{and} \quad \tau = o(np). \tag{D.10}
\]

Then
\[
\mathbb{P}\{ B \cap \mathcal{E} \} \leq n^{-1+o(1)}.
\]

**Proof.** Fix the root vertex \( u \) and two distinct vertices \( v, w \). We first upper bound \( \mathbb{P}\{ B_{u,v,w} \cap \mathcal{E} \} \).

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 \( N_w \) are at least \( n - 1 - 4np - (4np)^2 = n - o(n) \) in view of the assumption \( np = o(n^{1/2}) \). Therefore,
\[
\mathbb{P}\{ |\Pi_1(v)| \leq \tau, |\Pi_1(w)| \leq \tau \} \cap \mathcal{E} \mid A_{u,v} = 1, A_{u,w} = 1\}.
In view of Lemma 15 and the assumption that $\tau = o(np)$.
It follows that
\[
P \left\{ B_{u,v,w} \cap \mathcal{E} \right\} \leq p^2 \exp (-2(1 - o(1))np).
\]

By the union bound,
\[
P \left\{ B \cap \mathcal{E} \right\} \leq \sum_{u,v,w} P \left\{ B_{u,v,w} \cap \mathcal{E} \right\} \leq n^3p^2 \exp (-2(1 - o(1))np) \leq n^{-1+o(1)},
\]
where the last inequality holds due to $np \geq \log n$.

Now we are ready to prove our main proposition.

**Definition 4.** Let $H_u$ denote the event that tree $T(u)$ constructed by the graph matching process (3) satisfies

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

Define $H = \cap_u H_u$.

**Proposition 1.** Suppose (D.7) and (D.10) hold and $\tau \log n \geq C \log n$ for a sufficiently large universal constant $C > 0$. Then
\[
P \left\{ H \right\} \geq 1 - 3n^{-1+o(1)}.
\]

**Proof.** Note that
\[
(\cap_u A_u) \cap (B^c \cup E^c) \cap \mathcal{E} \subset H.
\]

Hence,
\[
P \left\{ H \right\} \geq 1 - \sum_u P \left\{ A_u^c \right\} - P \left\{ B \cap \mathcal{E} \right\} - P \left\{ E^c \right\}.
\]

In view of Lemma 15 and the assumption that $\tau \log n \geq C \log n$ for a sufficiently large universal constant $C > 0$, we have
\[
P \left\{ A_u^c \right\} \leq n^{-1}.
\]
By Lemma 16, we have

\[ P \{ B \cap E \} \leq n^{-1+o(1)}. \]

By Lemma 1, we have \( P \{ E^c \} \leq 1/n \). Then the desired conclusion readily follows.

**APPENDIX E  |  TIME COMPLEXITY OF ALGORITHM 1**

Recall that in Algorithm 1, we need to efficiently check whether there exists a set of \( m \) independent \( \ell \)-paths from a given vertex \( i_2 \) to a set of \( m \) seeded vertices \( L \subset \Gamma^G_\ell (i_2) \) in \( G_2 \) and a set of \( m \) independent \( \ell \)-paths from a given vertex \( i_1 \) to the corresponding seed set \( \pi_0(L) \subset \Gamma^G_\ell (i_1) \) 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 \( \mathcal{N}_\ell^G(i_1) \) of \( i_1 \) in \( G_1 \) up to radius \( \ell \). We delete all the edges \( (u, v) \) found if \( u \) and \( v \) 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 \( \tilde{\mathcal{N}}_\ell^G(i_1) \), with edges pointing away from \( i_1 \). Note that \( \tilde{\mathcal{N}}_\ell^G(i_1) \) may not be 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 \( \widetilde{\mathcal{N}}_\ell^G(i_2) \) denote the resulting local neighborhood of \( i_2 \).

Finally, we take the graph union of \( \tilde{\mathcal{N}}_\ell^G(i_1) \) and \( \tilde{\mathcal{N}}_\ell^G(i_2) \), by treating seeded vertex \( u \in \Gamma^G_\ell (i_2) \) with its correspondence \( \pi_0(u) \in \Gamma^G_\ell (i_1) \) as the same vertex. All the other vertices, seeded or nonseeded, from the two different local neighborhoods are treated as distinct vertices. We denote the resulting graph union as \( \mathcal{N}_\ell(i_1, i_2) \).

Recall that we aim to find a set of independent (vertex-disjoint except for \( i_1 \)) \( \ell \)-paths from \( i_1 \) to seeded vertices in \( \Gamma^G_\ell (i_1) \). 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 \( \mathcal{N}_\ell(i_1, i_2) \) into two vertices: \( u^+ \) and \( u^- \);
2. Add an edge of capacity 1 from \( u^+ \) to \( u^- \);
3. Replace each other edge \( (u, v) \) in \( \mathcal{N}_\ell(i_1, i_2) \) with an edge from \( u^- \) to \( v^+ \) of capacity 1;
4. Find a max-flow from \( i_1^- \) to \( i_2^+ \).

The idea behind this construction is as follows. Any flow path from the source vertex \( i_1^- \) to the sink vertex \( i_2^+ \) 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 [17]. 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 \( u^+ \) to \( u^- \), and the capacity here has been restricted to one. Also, the flow path must pass exactly 2\( \ell \) distinct \( u^- \) vertices (including the source vertex \( i_1^- \)), because all the edges are pointing away from \( i_1^- \) and towards \( i_2^+ \). Thus each flow path from \( i_1^- \) to \( i_2^+ \) represents a vertex-disjoint 2\( \ell \)-path from the source vertex \( i_1 \) to sink vertex \( i_2 \) in \( \mathcal{N}_\ell(i_1, i_2) \). As a consequence, the max-flow from \( i_1^- \) to \( i_2^+ \) corresponds to a maximal set of independent \( \ell \)-paths from \( i_2 \) to a set of seeded vertices \( L \subset \Gamma^G_\ell (i_2) \) in \( G_2 \), and of independent \( \ell \)-paths from \( i_1 \) to the corresponding seed set \( \pi_0(L) \subset \Gamma^G_\ell (i_1) \) in \( G_1 \).

As for time complexity, we can find a max-flow from \( i_1^- \) to \( i_2^+ \) via Ford-Fulkerson algorithm [17] in \( O(|E|f) \) time steps, where \(|E|\) is the total number of edges of \( \mathcal{N}_\ell(i_1, i_2) \) after vertex splitting and edge replacement, and \( f \) is the max flow. Under the choice of \( \ell \) given in (10), the total number of vertices and edges in \( \mathcal{N}_\ell(i_1, i_2) \) 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 $\Gamma^G(i_1)$ which is at most $O(n^{1/2 - \epsilon})$ with high probability. Hence, in total it takes $O(n\alpha)$ time steps to compute the max-flow from $i_1$ to $i_2^*$ via Ford-Fulkerson algorithm.

**APPENDIX F | TAIL BOUNDS FOR BINOMIAL DISTRIBUTIONS**

**Theorem 6 ([30, 33]).** Let $X \sim \text{Bin}(n, p)$. It holds that

\[
\mathbb{P}\{X \leq nt\} \leq \exp\left(-n\left(\sqrt{p} - \sqrt{t}\right)^2\right), \quad \forall 0 \leq t \leq p. \tag{F.1}
\]

\[
\mathbb{P}\{X \geq nt\} \leq \exp\left(-2n\left(\sqrt{t} - \sqrt{p}\right)^2\right), \quad \forall p \leq t \leq 1. \tag{F.2}
\]

\[
\mathbb{P}\{X \geq nt\} \leq 2^{-nt}, \quad \forall 6p \leq t \geq 1. \tag{F.3}
\]