The Power of D-hops in Matching Power-Law Graphs

LIREN YU, Purdue University, USA
JIAMING XU, Duke University, USA
XIAOJUN LIN, Purdue University, USA

This paper studies seeded graph matching for power-law graphs. Assume that two edge-correlated graphs are independently edge-sampled from a common parent graph with a power-law degree distribution. A set of correctly matched vertex-pairs is chosen at random and revealed as initial seeds. Our goal is to use the seeds to recover the remaining latent vertex correspondence between the two graphs. Departing from the existing approaches that focus on the use of high-degree seeds in 1-hop neighborhoods, we develop an efficient algorithm that exploits the low-degree seeds in suitably-defined D-hop neighborhoods. Specifically, we first match a set of vertex-pairs with appropriate degrees (which we refer to as the first slice) based on the number of low-degree seeds in their D-hop neighborhoods. This approach significantly reduces the number of initial seeds needed to trigger a cascading process to match the rest of graphs. Under the Chung-Lu random graph model with n vertices, max degree $\Theta(\sqrt{n})$, and the power-law exponent $2 < \beta < 3$, we show that as soon as $D > \frac{4-\beta}{3-\beta}$, by optimally choosing the first slice, with high probability our algorithm can correctly match a constant fraction of the true pairs without any error, provided with only $\Omega((\log n)^{4-\beta})$ initial seeds. Our result achieves an exponential reduction in the seed size requirement, as the best previously known result requires $n^{1/2+\epsilon}$ seeds (for any small constant $\epsilon > 0$). Performance evaluation with synthetic and real data further corroborates the improved performance of our algorithm.

CCS Concepts: • **Mathematics of computing → Random graphs**; • **Information systems → Data mining**.

Additional Key Words and Phrases: graph matching; power-law graphs; Chung-Lu model; multi-hop neighborhoods

ACM Reference Format:
Liren Yu, Jiaming Xu, and Xiaojun Lin. 2021. The Power of D-hops in Matching Power-Law Graphs. *Proc. ACM Meas. Anal. Comput. Syst.* 5, 2, Article 27 (June 2021), 43 pages. https://doi.org/10.1145/3410220.3460098

1 INTRODUCTION

Given two edge-correlated graphs, graph matching aims to find a bijective mapping between their vertex sets so that their edge sets are maximally aligned. It is a fundamental problem with numerous applications in a variety of fields, including social network de-anonymization [25], machine learning [11, 14], computer vision [10, 29], pattern recognition [4, 6], computational biology [17, 31] and natural language processing [16].

This paper focuses on seeded graph matching, wherein an initial set of seeds, i.e., correctly matched vertex-pairs, is revealed as side information. Seeded graph matching is motivated by the fact that in many real applications, some side information on the vertex identities is available,

Authors' addresses: Liren Yu, yu827@purdue.edu, Purdue University, 465 Northwestern Ave., West Lafayette, Indiana, USA, 47907-2035; Jiaming Xu, jx77@duke.edu, Duke University, 100 Fuqua Drive, Durham, North Carolina, USA, 27708; Xiaojun Lin, linx@purdue.edu, Purdue University, 465 Northwestern Ave., West Lafayette, Indiana, USA, 47907-2035.

2476-1249/2021/6-ART27. https://doi.org/10.1145/3410220.3460098

which has been successfully utilized to match many real-world networks [1] [24, 25]. Using seeds, we can then measure the similarity of a vertex-pair by its "witnesses." More precisely, let G_1 and G_2 denote two graphs. For each pair of vertices (u, v) with u in G_1 and v in G_2, a seed (w, w') is called a *1-hop witness* for (u, v) if w is a neighbor of u in G_1 and w' is a neighbor of v in G_2. Since G_1 and G_2 are graphs with correlated edges, a candidate pair of vertices are expected to have more witnesses if they are a true pair than if they are a fake pair. Not only has this idea been applied to many graph matching problems, strong performance guarantees (in term of the required number of seeds) have been obtained, in particular, for matching Erdős-Rényi graphs [15, 18, 19, 21–23, 27, 30, 33, 34].

However, Erdős-Rényi graphs fall short of capturing many fundamental structural properties of real-world networks. Notably, many real-world networks exhibit a power-law degree distribution, i.e., the fraction of nodes with degree at least k decays as $k^{-\beta+1}$ for some exponent $\beta > 0$. As a consequence, we expect to see very large degree fluctuations, with some nodes having very high degrees (so-called hubs) and some other sparsely-connected nodes with small degrees. Intuitively, this degree fluctuation may confuse witness-based vertex matching, e.g., a fake pair with high degrees may have many more witnesses than a true pair with low degrees, which foils the existing seeded algorithms designed for matching Erdős-Rényi graphs.

There have been several attempts to design seeded graph matching algorithms for power-law graphs [5, 7, 19]. However, they tend to require a larger number of seeds than Erdős-Rényi graphs. Note that to address the above-mentioned degree variations, a common idea is to first partition graphs into slices consisting of vertex-pairs with similar degrees. Then, the vertices are matched slice by slice, starting from the highest-degree slice to lower-degree slices. A cascade process is triggered, in the sense that the matched vertices in the current slice is used as new seeds to match the next slice. Intuitively, it is critical to correctly match the first slice in order to successfully trigger the cascading matching process for the later slices. [5, 7, 19] all use this idea and match the first slice based on 1-hop witnesses. Unfortunately, they also require a large number of correct seeds to match the first slice successfully. Specifically, [19] assumes preferential-attachment graphs with n vertices [2] and their algorithm requires $\Omega(n/\log(n))$ seeds to match a constant fraction of vertices correctly. [5, 7] instead assume the Chung-Lu graph model [8] (cf. Section 2). When all seeds are chosen from the high-degree vertices, [5, 7] show that their algorithm require only n^{ϵ} seeds to correctly match a constant fraction of the vertices. However, if the seeds are chosen uniformly from all vertices, the number of high-degree seeds will be much smaller than n^{ϵ}. In that case, the degree-driven graph matching (DDM) algorithm in [7] requires $n^{1/2+\epsilon}$ seeds to match a constant fraction of vertices correctly.

In this paper, we propose a new algorithm for matching power-law graphs. Our algorithm only requires $\Omega((\log n)^{4-\beta})$ initial seeds chosen randomly to correctly match a provably constant fraction of vertices. Our key departure from [5, 7, 19] is to use "witnesses" in larger D-hop neighborhoods. More precisely, a seed (w, w') is a D-hop witness for (u, v) if w is a D-hop neighbor of u in G_1 and w' is a D-hop neighbor of v in G_2. To see why using D-hop witnesses is crucial, note that, under the Chung-Lu model of [8] (cf. Section 2), even the highest degree vertices only have a 1-hop neighborhood of size at most $O(\sqrt{n})$. Since seeds are uniformly chosen, it is obvious that at least $\Omega(\sqrt{n})$ seeds are needed to ensure that a true pair in the first slice can have $\Omega(1)$ 1-hop witnesses. In contrast, as D increases, the size of the D-hop neighborhoods grows rapidly, and thus there are substantially more seeds that can serve as D-hop witnesses for true pairs, which provides hope to significantly reduce the number of initial seeds.

[1] For example, in social network de-anonymization, some users provide identifiable information in their service registrations or explicitly link their accounts across different social networks.

The idea of D-hop witnesses has also been used for matching Erdős-Rényi graphs in [23, 34]. However, as can be seen in the rest of the paper, the application of D-hop witnesses to power-law graphs is highly non-trivial. Specifically, due to the power-law degree variations, the D-hop neighborhoods of some high-degree vertices may become so large that even a fake pair can have many D-hop witnesses. Therefore, a key challenge is to properly control the size of the D-hop neighborhoods. This size depends not only on the degrees of the vertex-pairs to be matched, but also that of the intermediate nodes (to reach D-hop) and that of the seeds. To overcome this challenge, our algorithm design (to be explained in Section 3) (i) carefully chooses the first slice of vertices to be matched, (ii) carefully chooses the intermediate vertices when constructing the D-hop neighborhoods, and (iii) carefully avoids high-degree seeds in order to eliminate the confusion for fake pairs. These three ideas altogether ensure that the true pairs in the first slice have many more D-hop witnesses than the fake pairs, and thus can be correctly matched to trigger the cascading process to match the rest of the graphs. See Section 3 for more detailed discussions.

To fully realize the power of D-hops, we further need to carefully construct overlapping slices to account for the potential mismatch in the vertex slicing of graphs G_1 and G_2, and to design effective ways to match the remaining slices. Assembling all these pieces together enables us to achieve an exponential reduction in the required number of seeds compared to state-of-art results in [7]. Specifically, under the Chung-Lu model with power-law exponent $2 < \beta < 3$ and max degree $\Theta(\sqrt{n})$, we prove the following performance guarantee of our algorithm, stated informally here and formally in Section 5.

Theorem 1 (Summary of main result). *Suppose $D > \frac{4-\beta}{3-\beta}$. If there are $\Omega((\log n)^{4-\beta})$ initial seeds chosen independently at random, by optimally choosing the first slice, with high probability our algorithm correctly matches $\Omega(n)$ vertex-pairs without any error.*

This reduces the seed size requirement exponentially, as the best previously known result [7] requires $n^{1/2+\epsilon}$ seeds. To prove Theorem 1, there are several key innovations in our analysis in particular to address the difficult dependency issues across edges and slices. First, note that when we define the D-hop neighborhoods, we use vertex degrees to construct the slices and to select the seeds and intermediate nodes. This degree-based slicing unfortunately brings dependency issues. In particular, if we condition on the vertex degrees, then the edges are no longer independently generated according to the Chung-Lu model. To circumvent this dependency issue, we first show that the degree-guided construction and selection can be closely approximated by the weight-guided counterparts with high probability. Then we restore the independence by studying the weight-guided construction and selection, since the edges are independently generated according to the Chung-Lu model given the weights. Second, as we use the matched pairs in the current slice as new seeds to match the next slice, the matching results are correlated across different slices. To deal with these correlations, we carefully construct sets of matched pairs that only depend on vertex weights to "sandwich" the original set of matched pairs at each slice, but are not correlated any more, which allows us to eliminate the slice-dependency issue. Last but not least, to derive the optimal choice of the first slice and attain the smallest seed size requirement, we tightly bound the sizes of the common D-hop neighborhoods for both true pairs and fake pairs. Compared to the Erdős-Rényi graphs, this requires much more sophisticated lines of analysis of the neighborhood exploration process in the power-law graphs due to the heterogeneous vertex weights.

In the literature, the idea of D-hop witnesses has been used in Erdős-Rényi graphs [23, 34]. However, there is a significant difference in our results for power-law graphs. Specifically, in the Erdős-Rényi graphs with average degree d, the sizes of the D-hop neighborhoods are highly concentrated on d^D. Moreover, when the average degree d is a constant, the size of D-hop neighborhoods is always $O(1)$ for any constant D. Thus, unless D increases with n, at least $\Omega(n)$ seeds are still

needed to ensure that there are enough D-hop witnesses for true pairs. In stark contrast, the power of the D-hop becomes much more significant for matching power-law graphs. In particular, for power-law graphs with constant average degrees, by properly using the D-hop witnesses, we dramatically reduce the seed requirement to $\Omega((\log n)^{4-\beta})$, as soon as D exceeds $\frac{4-\beta}{3-\beta}$. Further, we note that the algorithms in [23, 34] do not need to worry about controlling the D-hop neighborhood, as they do not face the challenge of power-law degree variations.

Finally, we conduct extensive experiments on both synthetic and real power-law graphs to corroborate our theoretical analysis. In particular, we compare our algorithm with five other state-of-the-art seeded graph matching algorithms. Numerical results demonstrate that our algorithm drastically boosts the matching accuracy and requires substantially fewer seeds to correctly match a large fraction of vertices. Further, although our analysis focuses on matching two graphs of the same number of vertices, our algorithm can be readily applied to match two graphs of different sizes and return an accurate matching between vertices in the common subgraph of the two graphs. Indeed, our experiments on real networks in Section 6.3.2 and Section 6.3.3 show that our algorithm still achieves outstanding matching performance, even when two graphs are of very different sizes.

2 MODEL

Following [5, 7], we adopt the Chung-Lu random graph model [8] to generate the underlying parent graph with a power-law degree distribution. We use the Chung-Lu model because it can easily fit different power-law degree distributions of real graphs. Furthermore, when the Chung-Lu model is used to model large complex graphs, the small-world phenomenon, i.e., having short paths between any two vertices, is well captured with high probability [8]. The study in [28] also shows that the Chung-Lu model is effective to fit the eigenvalues and core decompositions for real graphs. Thus, we believe that our results under the Chung-Lu model can be applied to many real graphs as illustrated by our numerical experiments in Section 6. Here, $[n]$ denotes the set $\{1, 2, ..., n\}$.

Definition 1. Given parameters $\overline{w} > 0$, $\overline{w} \ll w_{\max} \le \sqrt{n\overline{w}}$, and $\beta > 2$, the Chung-Lu graph is a random graph $G_0([n], E)$ generated as follows. Each vertex $i \in [n]$ is associated with a positive weight $w_i = \overline{w}\frac{\beta-2}{\beta-1}\left(\frac{n}{i+i_0}\right)^{\frac{1}{\beta-1}}$, where $i_0 = n\left(\frac{\overline{w}(\beta-2)}{w_{\max}(\beta-1)}\right)^{\beta-1}$. For any pair of two vertices $i, j \in [n]$ with $i \neq j$, they are connected independently by an edge with probability $p_{ij} = \frac{w_i w_j}{n\overline{w}}$.

Note that i_0 is chosen such that $w_0 = w_{\max}$, which is the largest weight among all vertices. Further, $\overline{w}$ approximates the average weight as follows. Since $\overline{w} \ll w_{\max}$, it follows that $i_0 \ll n$. It can then be verified that $\frac{1}{n}\sum_{i=1}^{n} w_i \to \overline{w}$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{w_i \ge w\}} \propto w^{-\beta+1}$ as $n \to \infty$.[2] Thus, the degree of vertex i is expected to be close to w_i, which admits a power-law distribution with exponent β.

The Chung-Lu model is particularly convenient for modelling the degree variations in real-world networks. In these real-world networks, while the average degree is often a constant, a small but non-negligible fraction of the vertices has very large degrees (the so-called hubs) [3]. To model such sparse power-law graphs with hubs, we assume $\overline{w} = \Theta(1)$ and $2 < \beta < 3$. Empirical studies have shown that the vertex degrees of many real-world networks indeed follow a power-law distribution with $2 < \beta < 3$ [3, 9, 26]. Note that if $0 < \beta \le 2$, the average degree diverges and the network cannot be sparse; if $\beta \ge 3$, the degree variance is bounded and no large hub can appear [3].

[2] To see the first part of the statement, let $f(x) = w_x/n$. Then $\int_1^{n+1} f(x)dx \le \frac{1}{n}\sum_{i=1}^{n} w_i \le f(n) + \int_1^n f(x)dx$. Moreover, $\int_1^n f(x)dx = \overline{w}n^{\frac{2-\beta}{\beta-1}}\left((n+i_0+1)^{\frac{\beta-2}{\beta-1}} - (i_0+1)^{\frac{\beta-2}{\beta-1}}\right) \to w$, in view of $i_0 \ll n$ due to $w_{\max} \gg \overline{w}$. Further, we can verify the second part of the statement by $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{w_i \ge w\}} = \left(\frac{(\beta-2)\overline{w}}{(\beta-1)w}\right)^{\beta-1} - \frac{i_0}{n} \to \left(\frac{(\beta-2)\overline{w}}{(\beta-1)w}\right)^{\beta-1}$.

Next, we obtain a subgraph G_1 by sampling each edge of G_0 into G_1 independently with probability s, which is a constant independent of n. To construct another subgraph G_2, repeat the same sub-sampling process independently and relabel the vertices according to an *unknown* permutation $\pi : [n] \to [n]$. Throughout the paper, we denote a vertex-pair by (u, v), where $u \in G_1$ and $v \in G_2$. For each vertex-pair (u, v), if $v = \pi(u)$, then (u, v) is a true pair; if $v \neq \pi(u)$, then (u, v) is a fake pair.

Finally, there is an initial seed set $\mathcal{S}$ consisting of true pairs. Each true pair is added into $\mathcal{S}$ with probability θ independently. Our goal is to recover π based on the observation of G_1, G_2 and $\mathcal{S}$.

Notation. We use standard asymptotic notation: for two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$, if $a_n \leq Cb_n$ for some an absolute constant C and for all n; $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$, if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ or $a_n \asymp b_n$, if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$; $a_n = o(b_n)$ or $b_n = \omega(a_n)$, if $a_n/b_n \to 0$ as $n \to \infty$. For ease of reference, the key notations are summarized in Table 1. The definitions of the last seven notations will be provided in the sequel.

Table 1. Key Notations

G	Graph
n	Number of vertices
w_i	Vertex i's weight
$w_{\max}$	The largest vertex weight
β	Exponent of the power-law degree distribution
p_{ij}	Connection probability between vertices i and j in the parent graph
s	Sub-sampling probability
θ	Probability that a true pair is chosen as initial seed
$\mathcal{S}$	Seed set
$\Gamma_d^G(u)$	The set of d-hop neighbors of vertex u in graph G
P_k	The k-th perfect slice
Q_k	The k-th perfect slice-pair $Q_k \triangleq P_k \times P_k$
$\widehat{P}_k^G$	The k-th imperfect slice of graph G
$\widehat{Q}_k$	The k-th imperfect slice-pair $\widehat{Q}_k \triangleq \widehat{P}_k^{G_1} \times \widehat{P}_k^{G_2}$
n^γ	The largest weight of the first perfect slice
α_k	Threshold between the k-th perfect slice and the $(k-1)$-th perfect slice

3 KEY ALGORITHMIC IDEAS

In this section, we elaborate on our three design choices to properly control the D-hop neighborhood sizes: the weight of the seeds, the weight of the candidate vertex-pairs, and the weight of the intermediate vertices.

First, it is important to utilize low-weight seeds while avoiding high-weight seeds. Due to the power-law degree distribution, when seeds are uniformly chosen, there are many more low-weight seeds than high-weight seeds. Thus, the D-hop neighborhoods need to be large enough to reach sufficiently many low-weight seeds. However, for fake pairs, their large D-hop neighborhoods may also overlap. This implies that high-weight seeds may easily become witnesses for fake pairs, which can appear in many D-hop neighborhoods. Therefore, in order to avoid having too many witnesses for fake pairs, it is important to eliminate the high-weight seeds.

Second, for a given D, we need to carefully choose the first slice of candidate vertex-pairs to be matched using the D-hop witnesses. On the one hand, if the weight of the candidate vertex-pairs

is too small, the common D-hop neighborhoods of a true pair are too small to produce enough witnesses. On the other hand, if the weight of the candidate vertex-pairs is too large, the D-hop neighborhoods of a fake pair would intersect a lot, leading to too many D-hop witnesses.

Third, the high-weight vertices are not suitable to be the intermediate vertices in D-hop neighborhoods when D is large. This is because, when D is large, there exist some high-weight vertices with very large d-hop ($d < D$) neighborhoods. If these high-weight vertices become $(D-d)$-hop neighbors of the candidate vertices, the D-hop neighborhoods of the fake pairs would become too large. Thus, we should avoid using the high-weight vertices as the intermediate vertices.

Prompted by the above three ideas, we partition the graph into "perfect" slices

$$P_k = \{u : w_u \in [\alpha_k, \alpha_{k-1}]\} \quad \text{where } \alpha_k = n^\gamma/2^k \text{ for } k \geq 0, \text{ and } \alpha_{-1} = \infty, \tag{1}$$

for some $\gamma \in (0, \log_n w_{\max}]$. In particular, the first slice P_1 is the set of vertices with weight in $[n^\gamma/2, n^\gamma]$, which is the first set of the vertices that we wish to match. We will show in (8) that for a vertex in the first slice P_1, its number of $\Theta(1)$-weight D-hop neighbors is on the order of $n^{\gamma((3-\beta)(D-1)+1}$. Hence, we optimally choose γ close to $\frac{1}{(3-\beta)(D-1)+1}$ so that its number of $\Theta(1)$-weight D-hop neighbors is close to $\Theta(n)$. Under this optimal choice, we prove that sufficiently many vertex-pairs in the first slice are correctly matched so that they can be used as new seeds to trigger the cascading process to match the rest of the graphs slice-by-slice. In fact, for slice $k \geq 2$ until $k = k^*$ for some k^*, since the earlier slices provide so many new seeds, it turns out that using 1-hop witnesses suffices. When $k > k^*$, the slice-by-slice matching process stops, as there are not enough 1-hop witnesses to correctly match the slices with low-weight vertices. Fortunately, for the fake pairs with such low-weights, there are very few 1-hop witnesses as well. Thus we treat all the low-weight vertices as a single slice and apply the PGM algorithm in [33] to match them. Finally, we use all the matched vertex-pairs as new seeds to match the zero slice P_0 with very high weights.

For the above ideas to work, however, it is important that the earlier slices do not produce wrong matches; otherwise, the wrong matches will propagate errors to the subsequent slices. As such, we only match pairs with the number of witnesses larger than a threshold, as we will see next in the detailed algorithm.

4 THE POWER-LAW D-HOP (PLD) ALGORITHM

In this section, we present our Power-Law D-hop (PLD) algorithm, shown in Algorithm 1 and provide the intuition why it works. As we will explain below, a few steps of our PLD algorithm use the Greedy Maximum Weight Matching (GMWM) algorithm (which was also used in [1]) and the Percolation Graph Matching (PGM) algorithm (which was introduced in [33]). We will briefly explain GMWM and PGM below, and will provide their detailed description in Appendix A for reference.

4.1 Algorithm description

We first introduce some notations regarding D-hop neighborhoods. Given any graph G and two vertices u, v in G, we denote the length of the shortest path from u to v in G by $\text{dist}_G(u, v)$. For each vertex $u \in G$, the d-hop neighbors of u is denoted by $\Gamma_d^G(u) = \{v \in G : \text{dist}_G(u, v) = d\}$. The neighbors within d-hop of u is denoted by $N_d^G(u) = \bigcup_{j=1}^d \Gamma_j^G(u)$.

Our PLD algorithm carefully incorporates the key algorithmic ideas described in Section 3. At a high-level, we first slice the vertices according to their degrees. We then apply the D-hop algorithm to the first slice (which is carefully chosen). Afterwards, we apply the 1-hop algorithm to the lower-degree slices 2 to k^*, until the vertex degrees are about poly-logarithmic in n, in which case

we apply the PGM algorithm to the last slice with the lowest-degree vertices. Finally, we return to slice 0 of vertices with very high degrees.

The full algorithm is presented in Algorithm 1. We now describe the details.

Algorithm 1 The Power-Law D-hop (PLD) Algorithm.

1: **Input:** Graphs G_1 and G_2, initial seed set $\mathcal{S}$, parameters $D, \gamma, \tau_1, \tau_2, k^*$
2: Construct a subset of low-degree seeds $\widehat{\mathcal{S}} = \left\{(u,v) \in \mathcal{S} : \left|\Gamma_1^{G_1}(u)\right|, \left|\Gamma_1^{G_2}(v)\right| \le 5\log n\right\}$.
3: Let $\widehat{G}_i$ denote the subgraph of G_i induced by the vertex set $V_i = \left\{u : \left|\Gamma_1^{G_i}(u)\right| \le (1+\delta)n^\gamma s\right\}$ for $i = 1, 2$.
4: Partition the graph G_i into slices $\widehat{P}_k^{G_i}$ for $i = 1, 2$ and $0 \le k \le k^*$, according to (2).
5: In $\widehat{G}_1$ and $\widehat{G}_2$, for candidate vertex-pairs in $\widehat{Q}_1$, count their D-hop witnesses in $\widehat{\mathcal{S}}$ and use GMWM to match pairs with more than τ_1 D-hop witnesses (τ_1 is given in (3)). The set of matched pairs is $\mathcal{R}_1$.
6: **for** $k = 2$ to k^* **do**
7: For candidate vertex-pairs in $\widehat{Q}_k$, count their 1-hop witnesses in $\mathcal{R}_{k-1}$ and use GMWM to match pairs with more than $\tau_2(k)$ 1-hop witnesses ($\tau_2(k)$ is given in (5)). The set of matched pairs is $\mathcal{R}_k$.
8: **end for**
9: Let G_i' denote the subgraph of G_i induced by the vertex set $V_i' = \left\{u : \left|\Gamma_1^{G_i}(u)\right| \le (1+\delta)\alpha_{k^*-1}s\right\}$, for $i = 1, 2$.
10: Apply PGM to G_1' and G_2', with the seed set $\mathcal{R}_{k^*}$ and the threshold $r = 3$. The set of matched pairs is denoted by $\mathcal{R}_{k^*+1}$.
11: For candidate vertex-pairs in $\widehat{Q}_0$, count their 1-hop witnesses in $\widehat{\mathcal{R}} \triangleq \bigcup_{k=1}^{k^*+1} \mathcal{R}_k$ and match pairs with GMWM. The set of matched pairs is $\mathcal{R}_0$.
12: **Output:** All matched pairs $\mathcal{R} = \widehat{\mathcal{R}} \cup \mathcal{R}_0 \cup \mathcal{S}$

In line 2, we construct a subset of low-weight seeds to use as future witnesses. However, since we do not have access to the vertex weights directly, we construct a seed subset $\widehat{\mathcal{S}}$ that contains seeds with degrees no larger than $5\log n$ to ensure that all seeds with $\Theta(1)$ weights are included.

In line 3, we eliminate the vertices with degrees larger than $(1+\delta)n^\gamma$ and their adjacent edges, because we do not want to use the high-weight vertices as the intermediate vertices.

In line 4, we partition the graphs G_1 and G_2 into slices. Recall that the "perfect" slices P_k in (1) are defined with the vertex weights. Again, since we can not observe the vertex weight directly, we need to use the vertex degree as an estimate of the vertex weight. However, using vertex degree to slice vertices creates new technical difficulties. Specifically, for two vertices corresponding to a true pair, their actual degrees in G_1 and G_2 may differ, and thus these two vertices may be assigned to two slices of different indices in G_1 and G_2. This case becomes problematic because, if we only match slices with the same index, such a true pair would never be matched. Fortunately, the actual degrees of the vertices corresponding to a true pair should not differ too much (assuming a common sub-sampling probability s for both graphs). Thus, to address the above difficulty, we enlarge the slices a little bit, so that with high probability the two vertices corresponding to a true pair can fall into slices with the same index, and therefore have the opportunity to be matched. More precisely, for $k \ge 0$, we define the imperfect slice as

$$\widehat{P}_k^G = \left\{u : (1-\delta)\alpha_k s \le \left|\Gamma_1^G(u)\right| \le (1+\delta)\alpha_{k-1}s\right\}, \text{ for } k \ge 0, \tag{2}$$

where $\delta = \frac{1}{8}$ throughout this paper. Here, α_k are the same as (1), and the parameters γ and D will be set to satisfy (10) in Theorem 2. The imperfect slice-pair is then defined as $\widehat{Q}_k = \widehat{P}_k^{G_1} \times \widehat{P}_k^{G_2} = \{(u,v) : u \in \widehat{P}_k^{G_1}, v \in \widehat{P}_k^{G_2}\}$.

In line 5, we count the D-hop witnesses for all vertex-pairs in the first slices $\widehat{P}_1^{G_1}$ and $\widehat{P}_1^{G_2}$, and then use Greedy Maximum Weight Matching (GMWM) [1] to find the vertex correspondence such that the total number of witnesses is large. GMWM first finds the vertex-pair with the largest number of witnesses among all possible vertex-pairs. It then discards all vertex-pairs that are adjacent to the one just found, and chooses the vertex-pair with the largest number of witnesses among the remaining candidate vertex-pairs, and so on. The detailed description of the GMWM algorithm can be found in Appendix A.1. Here, we note that our earlier idea of enlarging the imperfect slices $\widehat{P}_k$ creates a new problem. That is, the imperfect slices with neighboring indices now have some overlap. As a result, it is possible that a slice pair contains a fake pair $(u, \pi(v))$, but does not contain the true pairs $(u, \pi(u))$ and $(v, \pi(v))$[3]. When that happens, the fake pairs $(u, \pi(v))$ may have the most witnesses among all the candidate vertex-pairs containing either u or $\pi(v)$. Thus, the fake pair $(u, \pi(v))$ may be matched by GMWM. Fortunately, the number of witnesses of these fake pairs is still expected to be smaller than that of any true pair. Therefore, to resolve this difficulty and to ensure that only the true pairs are matched, for the first slice we match only the vertex-pairs with no less than τ_1 D-hop witnesses, where τ_1 is set to be a constant fraction of the expected number of the D-hop witnesses for true pairs, i.e.,

$$\tau_1 = \frac{3}{10}\left(\frac{Cs^2}{12\overline{w}}\right)^D n^{\gamma((3-\beta)(D-1)+1)}\theta, \tag{3}$$

where $C \triangleq (2^{\beta-1}-1)\left(\frac{(\beta-2)\overline{w}}{(\beta-1)}\right)^{\beta-1}$. Similar thresholds are also used in the following steps when we match other slices.

In line 6-8, we use the matched pairs from the previous slice as new seeds, and use the 1-hop algorithm to match the vertices in slices $k = 2, ..., k^*$, where

$$k^* = \left\lfloor \log_2\left(n^\gamma\left(\frac{Cs^2}{192\overline{w}\log n}\right)^{\frac{1}{3-\beta}}\right)\right\rfloor. \tag{4}$$

In other words, we match the vertices with degrees larger than $(1-\delta)\alpha_{k^*}$, where $\alpha_{k^*} \geq \left(\frac{192\overline{w}\log n}{Cs^2}\right)^{\frac{1}{3-\beta}}$. Again to ensure that only the true pairs are matched for each slice, we only match the vertex-pairs with at least $\tau_2(k)$ 1-hop witnesses, where $\tau_2(k)$ is set to be half of the expected number of the 1-hop witnesses of the true pairs, i.e.,

$$\tau_2(k) = \frac{C\alpha_{k-1}^{3-\beta}s^2}{16\overline{w}}. \tag{5}$$

In line 9-10, we apply the PGM algorithm [33, Section 3], which iteratively matches vertex-pairs whose number of witnesses is no less than a threshold r (which is 3 in line 10), to match the remaining vertices with degrees no larger than $(1+\delta)\alpha_{k^*}$. Note that when the vertex weight is this small, estimating the vertex weight based on its degree is not accurate anymore. Thus, it is difficult to use the vertex degree to distinguish which slices should these vertices fall into. Instead, we treat all of these low-weight vertices as one slice. Further, for such low-degree vertices, using 1-hop

[3]This phenomenon does not contradict the idea of enlarging the slices. Enlarging the slices only guarantees the true pairs $(u, \pi(u))$ and $(v, \pi(v))$ are assigned into some slice-pairs. However, for other slice-pairs that contain the fake pair $(u, \pi(v))$, it is still possible that the two true pairs are not included.

algorithm based on the seeds from earlier slices will lead to poor performance, because even the true pairs in this slice have too few 1-hop witnesses. Fortunately, there are even fewer witnesses for the fake pairs with such low degrees. Thus, we can use the PGM algorithm. The PGM algorithm starts with the initial seed set $\mathcal{S}$ to calculate the number of witnesses for each vertex-pair. As soon as any vertex-pair gets at least r witnesses (which are called "marks"), it is matched and becomes a new seed. The "marks" of neighboring vertex-pairs are then updated to match more vertex-pairs. See Appendix A.2 for the description of the PGM algorithm. In this way, PGM can match a constant fraction of the rest of vertex-pairs, while avoiding matching fake pairs.

Finally, in line 11, the algorithm uses all vertex-pairs matched above as new seeds and matches the vertices in $\widehat{Q}_0$ via the 1-hop algorithm.

The total complexity of our algorithm is $O(n^{3-2\gamma(\beta-1)})$. The proof can be found in Appendix B.

4.2 Intuition

Before we present the main results, we explain the intuition why the above algorithm will work only with $\Omega((\log n)^{4-\beta})$ seeds. For the purpose of explaining this intuition, we ignore the inaccuracy of estimating the weights by the vertex degrees and assume that the graphs can be partitioned into perfect slices P_k. We further assume that the true mapping π is the identity permutation. Also, when we write $\approx$, we ignore the constant factors that are non-essential.

The key to the success of Algorithm 1 is appropriately choosing the first slice to apply the D-hop algorithm. We first calculate the probability that a vertex of $\Theta(1)$ weight lies in the D-hop neighborhood of a vertex in the first slice. Specifically, given a vertex u in the first slice P_1 and another vertex v of weight 1, we want to compute the probability q_D that v is a D-hop neighbor of u, i.e., $q_D \triangleq \mathbb{P}\left\{v \in \Gamma_D^{\widehat{G}_j}(u)\right\}$, where j is either 1 or 2. Note that if v is a D-hop neighbor of u, then v is connected to some $(D-1)$-hop neighbors i of u. Therefore, q_D satisfies the following recursion:

$$\begin{aligned} q_D &\approx \sum_{i\in\widehat{G}_j} \mathbb{P}\left\{v \in \Gamma_1^{\widehat{G}_j}(i)\right\} \times \mathbb{P}\left\{i \in \Gamma_{D-1}^{\widehat{G}_j}(u)\right\} \\ &\overset{(a)}{\approx} c\int_0^{n^\gamma} nw^{-\beta}\cdot\frac{w}{n\overline{w}}\cdot wq_{D-1}dw \\ &= c\frac{q_{D-1}}{\overline{w}}\int_0^{n^\gamma} w^{2-\beta}dw = \frac{cn^{\gamma(3-\beta)}}{\overline{w}(3-\beta)}q_{D-1}. \end{aligned} \tag{6}$$

In step (a), we integrate over the degree w of the $(D-1)$-hop neighbor i. Thus, $\mathbb{P}\left\{i \in \Gamma_{D-1}^{\widehat{G}_j}(u)\right\}$ is wP_{D-1} by our definition. Further, $w/(n\overline{w})$ is the probability that v (with weight 1) is connected to i, and number of such vertices i with degree in $[w, w+dw]$ is about $\sum_{i=1}^n \mathbf{1}_{\{w\le w_i\le w+dw\}} \to cnw^{-\beta}dw$ with $c = \left(\frac{(\beta-2)\overline{w}}{(\beta-1}\right)^{\beta-1}(\beta-1)$. By the Chung-Lu model, $q_1 \approx \frac{n^\gamma}{n\overline{w}}$. Iterating (6) over D, it follows that

$$q_D \approx \left(c\frac{n^{\gamma(3-\beta)}}{\overline{w}(3-\beta)}\right)^{D-1} q_1 \approx \frac{c^{D-1}n^{\gamma((3-\beta)(D-1)+1)}}{n\overline{w}^D(3-\beta)^{D-1}}. \tag{7}$$

As explained in Section 3, for the success of the D-hop algorithm, there are two key considerations. On the one hand, we need to ensure that the fake pairs in $Q_1 \triangleq P_1\times P_1$ have very few D-hop witnesses. As such, we want to prevent the fake pairs in Q_1 from having too many common neighbors of small weight. Therefore, we require $q_D \ll 1$ which roughly corresponds to $n^{\gamma((3-\beta)(D-1)+1)} \ll n$ and is close to the condition (10) (stated later in Theorem 2). On the other hand, we need to ensure

that the true pairs in Q_1 have sufficiently many $\Theta(1)$-weight D-hop witnesses. Indeed, for $u \in P_1$, its number of common D-hop neighbors of $\Theta(1)$-weight is at least

$$\left|\{v : w_v = \Theta(1)\} \cap \Gamma_D^{\widehat{G}_1 \wedge \widehat{G}_2}(u)\right| \approx nq_D \approx n^{\gamma((3-\beta)(D-1)+1)}, \tag{8}$$

where the first approximation holds because there are about $\Theta(n)$ vertices with $\Theta(1)$ weight based on the power-law weight distribution. Therefore, under condition (11) stated in Theorem 2, which is roughly $\theta = \Omega\left(\frac{\log n}{n^{\gamma((3-\beta)(D-1)+1)}}\right)$, all the true pairs have at least $\Omega(\log n)$ low-degree D-hop witnesses. The above choices thus ensure that all true pairs (but no fake pairs) are matched.

Interestingly, after matching the first slice, it triggers a cascading process, where the new matches at one slice can be used as new seeds to match the subsequent slice by the 1-hop algorithm. To see why using 1-hop witnesses is sufficient, recall that the weight of vertices in P_k satisfies

$$\alpha_k \le w_i \le \alpha_{k-1} \iff \frac{n}{\left(\frac{(\beta-1)\alpha_{k-1}}{(\beta-2)\overline{w}}\right)^{\beta-1}} - i_0 \le i \le \frac{n}{\left(\frac{(\beta-1)\alpha_k}{(\beta-2)\overline{w}}\right)^{\beta-1}} - i_0.$$

According to the index range of these vertices, we get that the number of vertices in P_k is $\Theta\left(n\alpha_{k-1}^{1-\beta}\right)$. Since the vertices in P_k and the vertices in P_{k+1} are connected independently with probability at least $\frac{\alpha_k\alpha_{k+1}}{n\overline{w}}$, it follows that, for a vertex in P_{k+1}, its number of 1-hop neighbors in P_k is about

$$n\alpha_{k-1}^{1-\beta} \times \frac{\alpha_k\alpha_{k+1}}{n\overline{w}} = \frac{\alpha_{k-1}^{1-\beta}\alpha_k\alpha_{k+1}}{\overline{w}} \ge \frac{\alpha_k^{3-\beta}}{8\overline{w}}. \tag{9}$$

Note that for the 1-hop algorithm to succeed, the true pairs need to have more than $\log n$ 1-hop witnesses [23]. Since $2 < \beta < 3$, we have $\frac{\alpha_k^{3-\beta}}{8\overline{w}} > \log n$, as long as $\alpha_k > \alpha_{k^*} \approx (\log n)^{\frac{1}{3-\beta}}$. Therefore, assuming that the true pairs in $Q_k \triangleq P_k \times P_k$ are correctly matched, we expect that the 1-hop algorithm can correctly match the true pairs in Q_{k+1} as long as $k < k^*$.

However, when $k \ge k^*$, for a vertex in P_{k+1}, its number of 1-hop neighbors in P_k becomes smaller than $\log n$, and thus the 1-hop algorithm can no longer match the vertices in P_{k+1} correctly. As discussed in Section 3, we instead resort to the PGM algorithm to match a constant fraction of the rest of low-weight vertices. Note that the key to the success of the PGM is that the number of witnesses for a fake pair is no more than 2 [33]. To see why this condition holds for the remaining low-weight vertices, note that the probability that a low-weight seed (with weight no larger than α_{k^*}) becomes a 1-hop witnesses for a fake pair with weight no larger than α_{k^*} is at most $\left(\frac{\alpha_{k^*}\alpha_{k^*}}{n\overline{w}}\right)^2 = \frac{\alpha_{k^*}^4}{n^2\overline{w}^2}$. Since there are at most n seeds and the majority of them are low-weight, the number of witnesses for any fake pair with low-weights is about $\frac{\alpha_{k^*}^4}{n\overline{w}^2} \lesssim \frac{(\log n)^{\frac{4}{3-\beta}}}{n\overline{w}^2} \ll 1$. Thus, we can use the PGM algorithm with threshold $r = 3$ to match a constant fraction of the low-weight vertex-pairs without errors.

Finally, the number of vertices with weight less than α_0 is $\Theta(n)$. If most true pairs with weight less than α_0 are matched, we can use them as new seeds to exactly match the remaining vertex-pairs in Q_0.

5 MAIN RESULTS

The following theorem provides a sufficient condition for our algorithm to correctly match a constant fraction of nodes without any errors. We define $C \triangleq (2^{\beta-1}-1)\left(\frac{(\beta-2)\overline{w}}{(\beta-1)}\right)^{\beta-1}$ and $\kappa \triangleq \frac{(1+2\delta)^2 2^{5-\beta} C}{(2^{3-\beta}-1)\overline{w}}$ throughout this paper.

Theorem 2. *Suppose $\gamma > 0$ and the positive integer D are chosen such that $\gamma \le \log_n w_{\max}$, $n^{2\gamma} = o(n)$, and*

$$n^{\gamma((3-\beta)(D-1)+1)} \le \frac{Cs(2^{3-\beta}-1)}{20\cdot 2^{3-\beta}} \left(\frac{Cs^2}{12\kappa^2\cdot\overline{w}}\right)^D \frac{n}{(\log n)^{3-\beta}}. \tag{10}$$

If the fraction θ of seeds satisfies

$$\theta \ge \frac{320\log n}{\left(\frac{Cs^2}{12\cdot\overline{w}}\right)^D n^{\gamma((3-\beta)(D-1)+1)}}, \tag{11}$$

then for all sufficiently large n, Algorithm 1 with τ_1 in (3) and $\tau_2(k)$ in (5) outputs $\Theta(n)$ true pairs and zero fake pairs with probability at least $1-n^{-1+o(1)}$.

Recall from (8) that $n^{\gamma((3-\beta)(D-1)+1)}$ is roughly the size of the D-hop neighborhood of a vertex (with weight around n^γ) in the first slice P_1. Therefore, on the one hand, (10) ensures that for two distinct vertices (u,v) in the first slice, the intersection of their D-hop neighborhoods is much smaller than the two neighborhoods, so that the fake pairs have much fewer D-hop witnesses than the true pairs. On the other hand, (11) ensures that the true pairs have at least $\Omega(\log n)$ D-hop witnesses.

Assuming $w_{\max} = \Theta(\sqrt{n})$, if we set $D=1$ and $\gamma = \frac{1}{2}-\epsilon$ for a small constant $\epsilon > 0$, then Theorem 2 recovers the seed requirement $n^{1/2+\epsilon}$ for the 1-hop algorithm which is comparable to the result in [7]. Surprisingly, for larger D, if we optimally choose n^γ in (13), then the seed requirement can be dramatically reduced to $\Omega((\log n)^{4-\beta})$, as shown by the following corollary.

Corollary 1 (The formal version of Theorem 1). *Suppose*

$$D \ge \frac{1}{3-\beta}\left(\frac{\log n}{\log(w_{\max})}-1\right)+1 \quad \text{and} \quad D > \frac{4-\beta}{3-\beta}. \tag{12}$$

Choose

$$n^{\gamma((3-\beta)(D-1)+1)} = \frac{cn}{(\log n)^{3-\beta}}, \tag{13}$$

for a sufficiently small constant c so that (10) is satisfied, and $\tau_1, \tau_2(k)$ according to (3) and (5), respectively. If the fraction of seeds satisfies

$$\theta \ge \frac{C_0\,(\log n)^{4-\beta}}{n}$$

for a sufficiently large constant C_0, then for all sufficiently large n, Algorithm 1 outputs $\Omega(n)$ true pairs and zero fake pairs, with probability at least $1-n^{-1}$.

According to (13), we choose γ asymptotically equal to $\frac{1}{[(3-\beta)(D-1)+1]}$. Condition (12) is imposed to ensure that this choice satisfies $\gamma < 1/2$ and $\gamma \le \log_n(w_{\max})$ in Theorem 2. Theorem 1 is a special case of Corollary 1, where $w_{\max} = \Theta(\sqrt{n})$ so that (12) reduces to $D > \frac{4-\beta}{3-\beta}$.

6 NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments to verify our theoretical findings and the effectiveness of the PLD algorithm. For all experimental results, we calculate the accuracy rate as the median of the proportion of vertices that are correctly matched, taken over 10 independent runs.

6.1 Choice of D and γ

In this section, we simulate our PLD algorithm with different D and γ to investigate the impact of the two parameters. We generate the underlying parent graph G_0 according to the Chung-Lu model with $n = 10000$, $\beta = 2.5$ and $\overline{w} = 10$. Then, we construct G_1 and G_2 by sampling each edge of G_0 twice independently with probability $s = 0.8$. The seeds are selected such that each true pair becomes a seed with probability θ independently.

In Fig. 1, we first plot the accuracy rates of our PLD algorithm with $D = 3$ and different γ, when θ varies from 0 to 0.01. We observe that for a given accuracy rate, when $\gamma = 1/[(3-\beta)(D-1)+1]$, the PLD algorithm requires the smallest number of seeds. This result is consistent with the theoretical prediction in Corollary 1, i.e., the optimal choice of γ approaches $1/[(3-\beta)(D-1)+1]$ as $n \to \infty$.

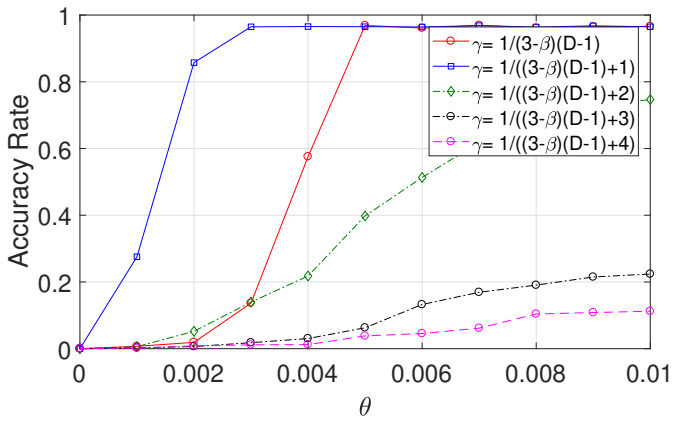

Fig. 1. The performance of the PLD algorithm with $D = 3$ and varying γ.

Then, in Fig. 2, we plot the accuracy rates of our PLD algorithm with different choices of D by fixing $\gamma = 1/[(3-\beta)(D-1)+1]$. We can see that the curves for different D align well with each other, showing that the PLD algorithm with different D requires a comparable number of seeds to succeed when γ is optimally chosen, as suggested by Corollary 1.

6.2 Performance Comparison with Synthetic Data

For our experiments on synthetic data, we still use the graphs generated in Section 6.1 according to the Chung-Lu model. Then, our PLD algorithm is simulated and compared with other five state-of-the-art seeded graph matching algorithms, namely DDM [7], Y-test [5], User-Matching [19], 2-hop [23] and PGM [33] algorithms. For the PLD algorithm, we select $D = 2, 3, 4$ and $\gamma = 1/((3-\beta)(D-1)+1)$ as suggested in Corollary 1. In Fig. 3, we plot the performance comparison when θ varies from 0 to 0.03. We observe that our PLD algorithm with different D achieves similar performance, and it significantly outperforms all other algorithms. Specifically, our PLD algorithm only requires around 50 seeds to match almost all vertices, while the User-Matching algorithm requires at least 150 seeds, and the DDM requires at least 220 seeds. Other algorithms perform even worse. Note that roughly 5% of vertices have degree at most 1 in both graphs; thus we do not expect to correctly match them. That is why the accuracy rates of our PLD algorithm saturated around 95%.

Note that the 2-hop and PGM algorithms have been known to work well for matching Erdős-Rényi graphs [23, 33]. However, we see that they are brittle to the power-law degree variations.

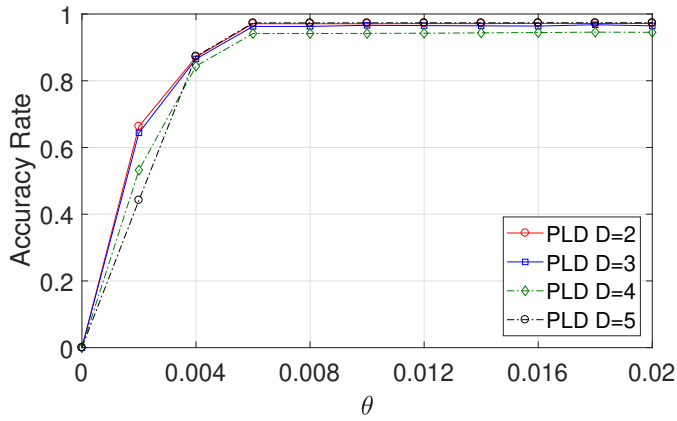

Fig. 2. The performance of the PLD algorithm with different D and $\gamma = \frac{1}{(3-\beta)(D-1)+1}$.

The DDM, Y-test, and User-Matching algorithms perform slightly better. However, since they all rely on the 1-hop witnesses, they still require a large number of seeds to succeed.

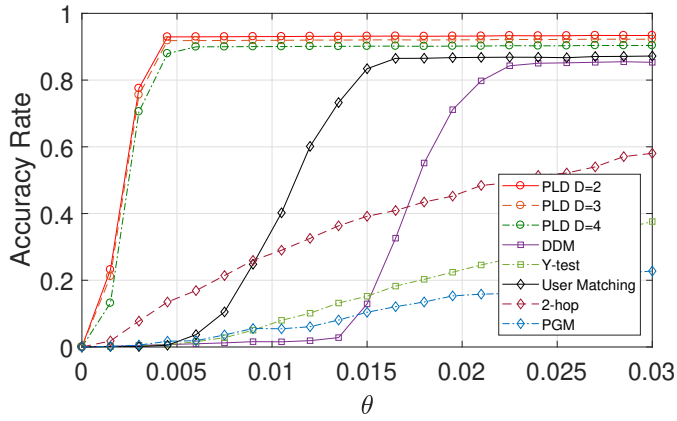

Fig. 3. Performance comparison of our PLD algorithm and five other algorithms on the Chung-Lu model with different θ.

In Fig. 4, we set $\theta = 0.01$ and plot the performance comparison when s varies from 0 to 1. We observe that, when s is small, the accuracy rates of all these algorithms are low. This is because a small correlation between the two graphs leads to insufficient witnesses for true pairs. For higher value of s, all algorithms' performance improves, but our PLD algorithm consistently outperforms other existing algorithms.

From Fig. 3 and Fig. 4, we can observe that the accuracy rate of the PLD algorithm exhibits a transition from nearly 0 to close-to-1, when θ goes from 0.0015 to 0.0045 in Fig. 3 and when s goes from 0.5 to 0.8 in Fig. 4. However, since our main result only provides a sufficient condition, we cannot conclude whether there is a sharp phase transition, which we leave for future work.

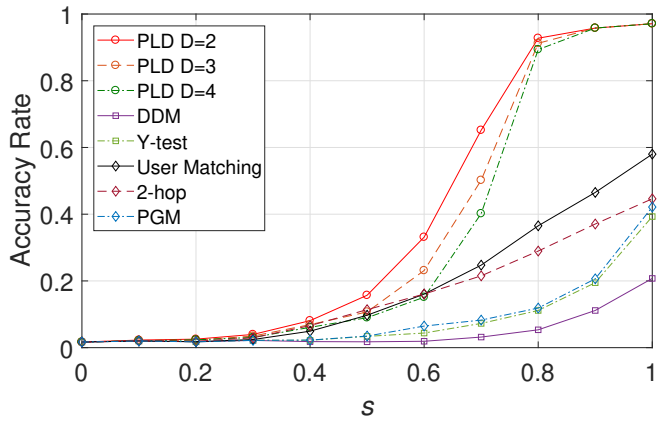

Fig. 4. Performance comparison of our PLD algorithm and five other algorithms on the Chung-Lu model with different s.

6.3 Performance Comparison with Real Data

6.3.1 Estimate Parameters for Real Graphs. We see that the performance of our PLD algorithm is outstanding on synthetic graphs. To further demonstrate the power of D-hops, we investigate its performance in matching real graphs. However, our algorithm based on the Chung-Lu model requires several parameters, which are unknown for real graphs. As such, in this section, we describe our method to estimate the key model parameters before implementing our algorithm.

First and foremost, we estimate the power-law exponent of real graphs by fitting them to the Chung-Lu model using the maximum-likelihood estimation given in [9]:

$$\widehat{\beta} = 1 + N \left[\sum_{d_i \geq d_{\min}} \ln \left(\frac{d_i}{d_{\min} - 1/2} \right) \right]^{-1}, \tag{14}$$

where d_i is the degree of vertex i, N is the number of vertices with degree at least $d_{\min}$, and $d_{\min}$ is some lower bound on the vertex degrees to be specified. It is suggested in [9] to estimate $d_{\min}$ using the Kolmogorov-Smirnov approach, which minimizes the maximum distance between the empirical CDF and the theoretical CDF of vertex degrees. More precisely,

$$d_{\min} = \arg\min_{d} \max_{d_i \geq d} \left| \widehat{F}_d(d_i) - F_d(d_i) \right|,$$

where $\widehat{F}_d(x)$ is the CDF of the observed vertex degrees with values at least d, and $F(x)$ is the CDF of the power-law vertex distribution restricted to $[d, +\infty)$. Numerical experiments in [9] show $\widehat{\beta}$ is accurate to 1% or better if $d_{\min}$ is set to be around 6. Thus, we fix $d_{\min} = 6$ throughout our real-data experiments.

Next, we estimate the subsampling probability s, which characterizes the edge correlation between the two observed graphs. Let $G_j[S]$ denote the subgraph of G_j induced by vertices in $S = \{i : (i, i) \in \mathcal{S}\}$, where $\mathcal{S}$ is the initial seed set. Note that under our subsampling model, given an edge in one graph, it appears in the other graph with probability s. Thus we estimate the sampling probability s by

$$\widehat{s} = \frac{2\,|E[G_1[S] \wedge G_2[S]]|}{|E[G_1[S]]| + |E[G_2[S]]|}, \tag{15}$$

where $E[G]$ denotes the edge set of graph G.

Based on $\widehat{s}$, we can further estimate the average weight $\overline{w}$. Recall that $\overline{w}$ is close to the average degree under the Chung-Lu model. Thus, we estimate $\overline{w}$ by $\frac{\overline{d}(G_1)+\overline{d}(G_2)}{2\widehat{s}}$, where $\overline{d}(G)$ is the average degree in graph G. Finally, for the fraction of seeds θ, if it is unknown, we can simply estimate it by $\frac{|\mathcal{S}|}{n}$. Note that since $w_{\max}$ will not be used by our algorithm, we do not need to estimate it.

Based on the estimated model parameters, we can then determine the input parameters of our PLD algorithm. Since we optimally choose $\gamma = 1/((3-\beta)(D-1)+1)$, the threshold τ_1 in (3) can be simplified to $\tau_1 = \frac{3}{10}\left(\frac{Cs^2}{12\overline{w}}\right)^D n\theta$. Further, the threshold $\tau_2(k)$ can be set according to (5).

6.3.2 Facebook Friendship Networks. We use a Facebook friendship network (provided in [32]) of 63392 students and staffs from University of Oregon as the parent graph G_0. There are 1633772 edges in G_0. The power-law exponent of the Facebook social network is estimated as 2.09 by (14). To obtain two edge-correlated subgraphs G_1 and G_2 of different sizes, we independently sample each edge of G_0 twice with probability $s = 0.9$ and sample each vertex of G_0 twice with probability 0.8. Then, we relabel the vertices in G_2 according to a random permutation $\pi : [n_2] \to [n_2]$, where n_2 is the number of nodes in G_2. Let m denote the number of common vertices that appear in both G_1 and G_2. The initial seed set is constructed by including each true pair independently with probability θ. We treat G_1 as the public network and G_2 as the private network, and the goal is to de-anonymize the node identities in G_2 by matching G_1 and G_2. In Fig. 5, we show the performance of our PLD algorithm and five other algorithms, when the fraction of initial seeds θ varies from 0 to 0.05. We can observe that our PLD algorithm significantly outperforms other algorithms. To investigate which types of vertices contribute most to the matching error of our PLD algorithm, we fix $\theta = 0.01$ and plot in Fig. 6 the statistics of the wrongly matched vertices according to their degrees. Specifically, for a given degree, we compute the number of vertices with such a degree in the parent graph that are not correctly matched by the PLD algorithm. We can observe that most matching errors are from low-degree vertices. This is because the low-degree vertices do not have sufficient common neighbors to distinguish the true pairs. Since there are about 23.5% vertices that have degree at most 1 in either G_1 or G_2 (who are thus difficult to be correctly matched), the matching accuracy in Fig. 5 is saturated at around 75%.

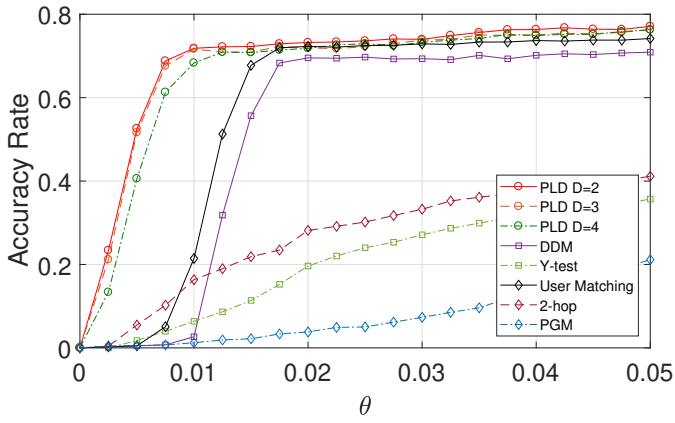

Fig. 5. Performance comparison of the PLD algorithm and five other algorithms applied to the Facebook networks with different θ.

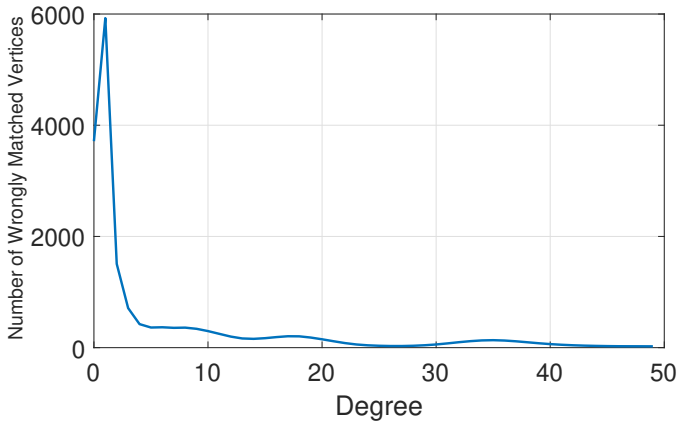

Fig. 6. The statistics of the wrongly matched pairs according to their degrees (in the parent graphs) when the PLD algorithm is applied to the Facebook networks.

In Fig. 7, we set $\theta = 0.01$ and plot the performance comparison when s varies from 0 to 1. Similar to Fig. 4, for small s, the accuracy rates of all matching algorithms become small because the correlation between the two graphs are low. For larger s, our PLD algorithm again consistently outperforms other algorithms.

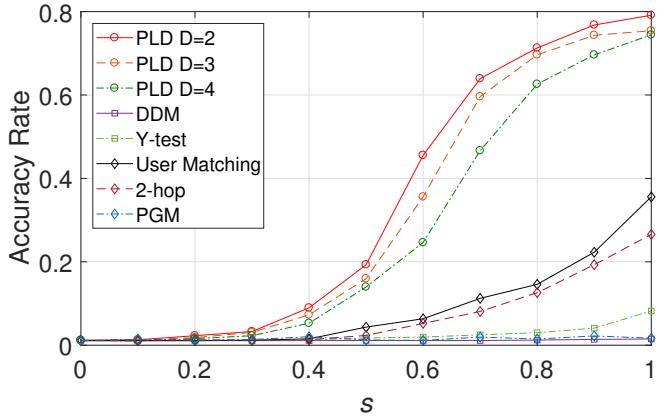

Fig. 7. Performance comparison of the PLD algorithm and five other algorithms applied to the Facebook networks with different s.

6.3.3 Autonomous Systems Networks. Following [13], we use the Autonomous Systems (AS) data set from [20] to further test the graph matching performance on power-law graphs. The data set consists of 9 graphs of Autonomous Systems peering information inferred from Oregon route-views between March 31, 2001, and May 26, 2001. Since some vertices and edges are changed over time, these nine graphs can be viewed as correlated versions of each other. The number of vertices of the 9 graphs ranges from 10,670 to 11,174 and the number of edges from 22,002 to 23,409. We aim to

match each graph to that on March 31, with vertices randomly permuted. The initial seed set is obtained by including each true pair independently with probability θ.

The power-law exponent of the Autonomous Systems networks is estimated to be 2.01 according to (14). Note that in this experiment, the two correlated graphs are provided by the real data set. Thus, we further estimate the correlation parameter s according to (15).

The performance comparison of the six algorithms is plotted in Fig. 8 for $\theta = 0.1$ and in Fig. 9 for $\theta = 0.01$. We observe that our PLD algorithm again significantly outperforms other algorithms. Note that the accuracy rates for all algorithms decay in time, because over time the graphs become less correlated with the initial one on March 31.

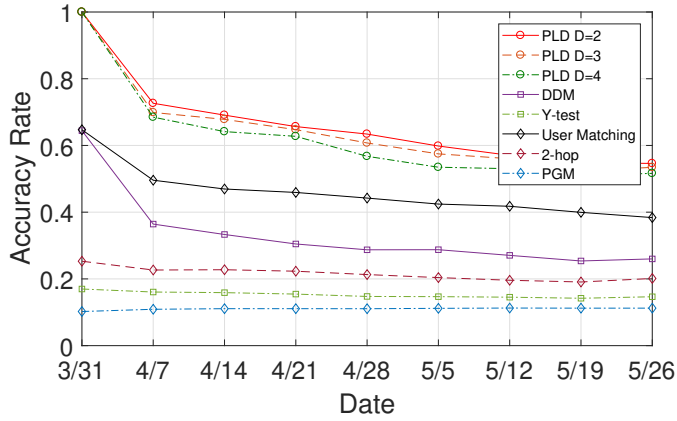

Fig. 8. Performance comparison of the PLD algorithm and five other algorithms applied to the Autonomous Systems graphs when $\theta = 0.1$.

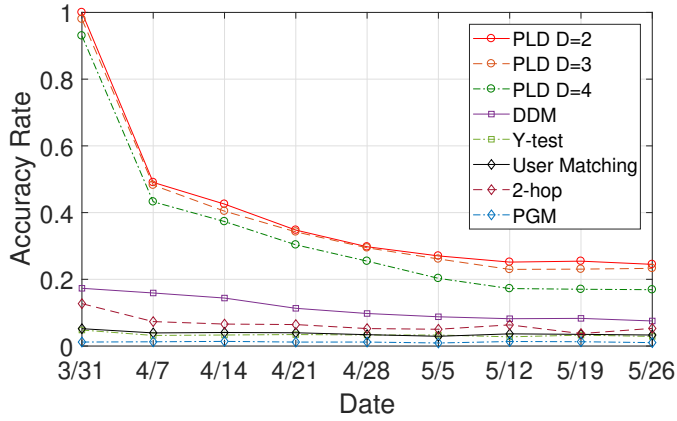

Fig. 9. Performance comparison of the PLD algorithm and five other algorithms applied to the Autonomous Systems graphs when $\theta = 0.01$.

7 ANALYSIS

In this section, we present the proof for Theorem 2. In Section 7.1, we describe the dependency issue in our analysis and how we deal with it. In Section 7.2, we prove that all the true pairs in the first slice Q_1 are matched error-free by the D-hop algorithm. Using the matched vertices in the previous slice as new seeds, we show in Section 7.3 that all the true pairs in slice Q_k are matched error-free by the 1-hop algorithm for $2 \le k \le k^*$. Further, Section 7.4 proves that using the match pairs in slice k^* as new seeds, the PGM algorithm correctly matches a constant fraction of true pairs with low weights. Finally, in Section 7.5, we come back to Q_0 and prove that using all the matched pairs as seeds, all the true pairs in Q_0 are matched error-free by the 1-hop algorithm. Theorem 2 readily follows by combining these results. The proofs of auxiliary lemmas can be found in Appendix C.2.

For ease of presentation, throughout the analysis, we assume without loss of generality that the true mapping π is the identity permutation. We further assume $\gamma > 0$ and the integer D are such that $\gamma \le \log_n w_{\max}$, $n^{2\gamma} = o(n)$, and (10) holds.

7.1 Deal with the Dependency Issues

In Algorithm 1, we use degrees as guidance to define the imperfect slice $\widehat{P}_k^{G_j}$ for $j = 1, 2$ and the induced graphs $\widehat{G}_1, \widehat{G}_2$. However, if we condition on the degrees, then the edges are no longer independently generated with probability p_{ij} as defined in the Chung-Lu model. To deal with this dependency issue, we construct slices based on vertex weight that "sandwich" $\widehat{P}_k^{G_j}$. Recall that the perfect slices defined as $P_k = \{u : w_u \in [\alpha_k, \alpha_{k-1}]\}$. By construction and the concentration of vertex degrees, we expect that $P_k \subset \widehat{P}_k^{G_j}$. We also need another weight-guided slice to contain $\widehat{P}_k^{G_j}$. Specifically, define

$$\overline{P}_k = \{u : w_u \in [(1-2\delta)\alpha_k, (1+2\delta)\alpha_{k-1}]\},$$

where $\delta = \frac{1}{8}$. We also define $\overline{Q}_k \triangleq \overline{P}_k \times \overline{P}_k$. The following lemma shows that with high probability, $P_k \subset \widehat{P}_k^{G_j} \subset \overline{P}_k$ and hence $Q_k \subset \widehat{Q}_k \subset \overline{Q}_k$. Similarly, we define two different subsets of vertices that "sandwich" V_j:

$$\underline{V} = \{u : w_u \in [0, n^\gamma]\} \quad \text{and} \quad \overline{V} = \{u : w_u \in [0, (1+2\delta)n^\gamma]\}.$$

Further, let $\underline{G}_j$ and $\overline{G}_j$ denote the subgraph of G_j induced by the vertex set $\underline{V}$ and $\overline{V}$, respectively, for $j = 1, 2$. The following lemma shows that with high probability, $\underline{V} \subset V_j \subset \overline{V}$ and hence $\underline{G}_j \subset \widehat{G}_j \subset \overline{G}_j$.

Lemma 1. *For any $0 \le k \le k^*$,*

$$\mathbb{P}\left\{Q_k \subset \widehat{Q}_k \subset \overline{Q}_k\right\} \ge 1 - n^{-4+o(1)},$$

and

$$\mathbb{P}\left\{Q_{\ge k^*} \subset \widehat{Q}_{\ge k^*} \subset \overline{Q}_{\ge k^*}\right\} \ge 1 - n^{-4+o(1)}.$$

For $j = 1, 2$,

$$\mathbb{P}\left\{\underline{V} \subset V_j \subset \overline{V}\right\} = \mathbb{P}\left\{\underline{G}_j \subset \widehat{G}_j \subset \overline{G}_j\right\} \ge 1 - n^{-3+o(1)}.$$

7.2 Match Pairs in $\widehat{Q}_1$ using D-hop Algorithm

Recall that we give a heuristic argument of (8), showing that for a true pair in Q_1, the number of common D-hop neighbors of $\Theta(1)$ weights is on the order of $n^{\gamma(3-\beta)(D-1)+1}$, by ignoring the the potential dependency between $\widehat{G}_j, \widehat{Q}_1$ and graphs G_1, G_2. To resolve this dependency, we crucially

exploit the fact that with high probability $Q_1 \subset \widehat{Q}_1$ and $\underline{G}_j \subset \widehat{G}_j$ as shown in Lemma 1. In particular, we consider a true pair (u,u) in Q_1 and bound its number of $\Theta(1)$-weight D-hop neighbors in $\underline{G}_j$. Unfortunately, even when $\underline{G}_j \subset \widehat{G}_j$, the D-hop neighbors of u in $\underline{G}_j$ may contain some vertices that are within the $(D-1)$-hop neighborhood of u in $\widehat{G}_j$, which means $\Gamma_D^{\underline{G}_j} \not\subseteq \Gamma_D^{\widehat{G}_j}$. In order to exclude such vertices, we bound the number of $\Theta(1)$-weight vertices in $N_{D-1}^{\overline{G}_j}(u)$ from above. Fortunately, $\left|N_{D-1}^{\overline{G}_j}(u)\right|$ is close to $\left|\Gamma_{D-1}^{\overline{G}_j}(u)\right|$, which is on the order of $n^{\gamma(3-\beta)(D-2)+1}$ and thus is much smaller than $\left|\Gamma_D^{\underline{G}_j}\right|$. To be more precise, we have the following lemma.

Lemma 2. *Fix any vertex $u \in P_1$ and constant c. For all sufficiently large n,*

$$\mathbb{P}\left\{\left|\Gamma_D^{\underline{G}_1 \wedge \underline{G}_2}(u) \cap \{i : w_i \le c\}\right| \ge \Gamma_{\min}\right\} \ge 1 - n^{-4+o(1)}, \tag{16}$$

$$\mathbb{P}\left\{\left|N_{D-1}^{\overline{G}_j}(u) \cap \{i : w_i \le c\}\right| \le N_{\max}\right\} \ge 1 - n^{-4+o(1)}, \text{ for } j = 1,2, \tag{17}$$

where $\Gamma_{\min} = \frac{1}{2}\left(\frac{C \cdot s^2}{12 \cdot \bar{w}}\right)^D n^{\gamma((3-\beta)(D-1)+1)}$ *and* $N_{\max} = 2c\kappa^D n^{\gamma((3-\beta)(D-2)+1)}$.

To appreciate the utility of Lemma 2, note that under the high-probability event $\underline{G}_j \subset \widehat{G}_j \subset \overline{G}_j$ for $j = 1, 2$, we have

$$\Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(u) \supset \Gamma_D^{\underline{G}_1 \wedge \underline{G}_2}(u) \setminus \left(N_{D-1}^{\overline{G}_1}(u) \cup N_{D-1}^{\overline{G}_2}(u)\right).$$

Therefore, combining (16) and (17) implies that with high probability,

$$\left|\Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(u) \cap \{i : w_i \le c\}\right| \ge \Gamma_{\min} - 2N_{\max} \approx \Gamma_{\min}, \tag{18}$$

where the last approximation holds because $\Gamma_{\min} \gg N_{\max}$ due to $2 < \beta < 3$. Hence, the last display yields the desired lower bound (8) to the number of common D-hop neighbors of $\Theta(1)$ weights for a true pair (u,u) in Q_1.

Next, we adopt a similar strategy to study fake pairs. In particular, for a fake pair in $\widehat{Q}_1$, we bound from above its number of common D-hop neighbors of weights smaller than $\frac{15}{s}\log n$.[4] Again, to circumvent the dependency between $\widehat{G}_j, \widehat{Q}_1$ and graphs G_1, G_2, we consider a fake pair (u,v) in $\overline{Q}_1$ and bound from above its number of $\Theta(1)$-weight neighbors within the common D-hop neighborhood in $\overline{G}_1$ and $\overline{G}_2$.

Lemma 3. *Fix any two distinct vertices $u, v \in \overline{P}_1$. For sufficiently large n,*

$$\mathbb{P}\left\{\left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \cap \{i : w_i \le \frac{15}{s}\log n\}\right| \le \Psi_{\max}\right\} \ge 1 - n^{-4+o(1)}, \tag{19}$$

where $\Psi_{\max} = \frac{2^{3-\beta}\kappa^{2D} n^{2\gamma((3-\beta)(D-1)+1)}}{(2^{3-\beta}-1)Cn}\left(\frac{15}{s}\log n\right)^{3-\beta} + \frac{2^{\beta-2}}{2^{\beta-2}-1}\kappa^{D-1} n^{(\gamma(3-\beta)(D-2)+1)}(4 + 6\log n)$.

Remark 1. To see how (19) follows, note that

$$N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \subset \left(\Gamma_D^{\overline{G}_1}(u) \cup N_{D-1}(u,v)\right) \cap \left(\Gamma_D^{\overline{G}_2}(v) \cup N_{D-1}(u,v)\right) = \left(\Gamma_D^{\overline{G}_1}(u) \cap \Gamma_D^{\overline{G}_2}(v)\right) \cup N_{D-1}(u,v),$$

where $N_{D-1}(u,v) = N_{D-1}^{\overline{G}_1}(u) \cup N_{D-1}^{\overline{G}_2}(v)$. We have already obtained an upper bound to $\left|N_{D-1}^{\overline{G}_j}\right|$ when proving (17) for $j = 1, 2$. Thus, it remains to bound from above $\left|\Gamma_D^{\overline{G}_1}(u) \cap \Gamma_D^{\overline{G}_2}(v)\right|$. A simple yet key

[4] The threshold $\frac{15}{s}\log n$ is chosen such that $\{i : w_i \le \frac{15}{s}\log n\}$ contains $\{i : |\Gamma_1^{G_1}(i)| \le 5\log n, |\Gamma_1^{G_2}(i)| \le 5\log n\}$ with high probability.

observation is that for a vertex i of weight 1, there are two extreme cases in which i becomes a common D-hop neighbor of (u,v). One case is that i connects to some vertex in $\Gamma_{D-1}^{\overline{G}_1}(u) \setminus \Gamma_{D-1}^{\overline{G}_2}(v)$, and connects to some other vertex in $\Gamma_{D-1}^{\overline{G}_2}(v) \setminus \Gamma_{D-1}^{\overline{G}_1}(u)$. It can be shown that each of these two connections happens independently with probability approximately q_D and thus the number of such common D-hop neighbors is about nq_D^2, which roughly gives rise to the first term of $\Psi_{\max}$. The other extreme case is that i is a $(D-1)$-hop neighbor of some common neighbor of (u,v). Luckily, the common 1-hop neighborhood of (u,v) is typically of a very small size and thus we can bound from above $\left|\Gamma_1^{\overline{G}_1}(u) \cap \Gamma_1^{\overline{G}_2}(v)\right|$ by approximately $\log n$. Moreover, i becomes a $(D-1)$-hop neighbor of a given vertex in $\Gamma_1^{\overline{G}_1}(u) \cap \Gamma_1^{\overline{G}_2}(v)$ with probability at most q_{D-1}. Thus, the number of such common D-hop neighbors is at most around $nq_{D-1}\log n$, which gives an expression close to the second term of $\Psi_{\max}$. These two extreme cases turn out to be the dominating cases as shown in the proof of Lemma 3.

To see the usage of Lemma 3, note that under the high-probability event $\widehat{G}_j \subset \overline{G}_j$ for $j=1,2$, we have $\Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(v) \subset N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v)$. Therefore, (19) implies that with high probability

$$\left|\Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(v) \cap \{i : w_i \le \frac{15}{s}\log n\}\right| \le 2\Psi_{\max}, \tag{20}$$

which yields the desired upper bound to the number of common D-hop neighbors of $\Theta(1)$ weights for a fake pair (u,v) in $\widehat{Q}_1$.

Finally, since we have $n^{\gamma(3-\beta)} \gg \log n$ and $n^{\gamma((3-\beta)(D-1)+1)}(\log n)^{3-\beta} = O(n)$ based on the choice in (11), it follows that $\Gamma_{\min} > 2\Psi_{\max}$. Moreover, (11) ensures that $\Gamma_{\min}\theta = \Omega(\log n)$. Therefore, combining (18) and (20) implies that the true pairs in Q_1 have more D-hop witnesses than the fake pairs in $\widehat{Q}_1$. Hence, we can use Algorithm 1 to match pairs in $\widehat{Q}_1$ correctly. More precisely, we have the following lemma.

Lemma 4. *Under the conditions of Theorem 2, for all sufficiently large n, the set of matched pairs in Step 5 of Algorithm 1, denoted by $\mathcal{R}_1$, contains all true pairs in Q_1 and no fake pairs in $\widehat{Q}_1$ with probability at least $1 - n^{-1.5+o(1)}$.*

7.3 Match Pairs in $\widehat{Q}_k$ Slice by Slice using 1-hop Algorithm

Given that all the true pairs in Q_1 are matched error-free, we show that all the true pairs in Q_k are matched error-free by the 1-hop algorithm for all $2 \le k \le k^*$.

Note that when matching pairs in $\widehat{Q}_k$, we use $\mathcal{R}_{k-1}$, the set of matched vertices in $\widehat{Q}_{k-1}$, as seeds. Suppose slice $k-1$ is successfully matched. Then, $\mathcal{R}_{k-1}$ contains all the true pairs in Q_{k-1}. Therefore, for a true pair in Q_k, to bound from below its number of 1-hop witnesses in $\mathcal{R}_{k-1}$, it suffices to consider its number of 1-hop common neighbors in P_{k-1}, which is on the order of $\frac{\alpha_{k-1}^{3-\beta}}{2\overline{w}}$ as we explained in (9). This intuition is made precise by the following lemma.

Lemma 5. *Fix any $2 \le k \le k^*$ and any vertex $u \in P_k$. For all sufficiently large n,*

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k-1}\right| \ge \xi_k\right\} \ge 1 - n^{-4}, \tag{21}$$

where $\xi_k = \frac{C\alpha_{k-1}^{3-\beta}s^2}{16\overline{w}}$.

Moreover, if slice $k-1$ is successfully matched, since there is no matching error, on the high-probability event $\widehat{P}_{k-1} \subset \overline{P}_{k-1}$, $\mathcal{R}_{k-1}$ is contained by the set of true pairs in $\overline{Q}_{k-1} \triangleq \overline{P}_{k-1} \times \overline{P}_{k-1}$. Therefore, for a fake pair in $\widehat{Q}_k$, to bound from above its the number of 1-hop witnesses in $\mathcal{R}_{k-1}$, it

suffices to bound its number of 1-hop common neighbors in $\overline{P}_{k-1}$, which is done in the following lemma. Note that to resolve the potential dependency between $\widehat{Q}_k$ and graphs G_1, G_2, we state the lemma for a fake pair in $\overline{Q}_k$, which contains $\widehat{Q}_k$ with high probability.

Lemma 6. *Fix any $2 \le k \le k^*$ and any two distinct vertices $u, v \in \overline{P}_k$, Then for all sufficiently large n,*

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{k-1}\right| \le \zeta_k\right\} \ge 1 - n^{-4}, \tag{22}$$

where $\zeta_k = \frac{8(1+2\delta)^4 C \alpha_{k-1}^{5-\beta}}{\overline{w}^2 n} + \frac{16}{3}\log n$.

To see how (22) follows, note that a vertex in $\overline{P}_{k-1}$ is a 1-hop common neighbor for the fake pair (u, v) with probability at most on the order of $\left(\frac{\alpha_k \alpha_{k-1}}{n\overline{w}}\right)^2 = \frac{\alpha_{k-1}^4}{4n^2\overline{w}^2}$. Since there are $\Theta(n\alpha_{k-1}^{1-\beta})$ vertices in $\overline{P}_{k-1}$, the number of 1-hop common neighbors in $\overline{P}_{k-1}$ is about $\frac{\alpha_{k-1}^{5-\beta}}{4n\overline{w}^2}$ on expectation. The extra term $\frac{16}{3}\log n$ in (22) comes from the sub-exponential tail bounds when we apply concentration inequalities.

Recall that we assume $n^{2\gamma} = o(n)$ and hence $\alpha_{k-1}^{3-\beta} \gg \frac{\alpha_{k-1}^{5-\beta}}{n}$ for $2 \le k \le k^*$. Moreover, $\alpha_{k-1}^{3-\beta} \ge \alpha_{k^*}^{3-\beta} \ge \frac{192\overline{w}\log n}{Cs^2}$ for $2 \le k \le k^*$. It then can be verified that $\xi_k > \zeta_k$. Thus, we expect that the 1-hop algorithm can match vertex-pairs in $\widehat{Q}_k$ correctly. More precisely, we have the following lemma.

Lemma 7. *Under the conditions of Theorem 2, for all sufficiently large n, with probability at least $1 - n^{-1.5+o(1)}$, the set of matched pairs in Step 6-8 of Algorithm 1, denoted by $\mathcal{R}_k$, contains all true pairs in Q_k and no fake pairs in $\widehat{Q}_k$ for all $2 \le k \le k^*$.*

7.4 Match Low-Weight Pairs by PGM

We proceed to match pairs with weight smaller than α_{k^*} using the PGM algorithm. As explained in Section 4.2, we expect that the number of common 1-hop neighbors for any fake pair with weights smaller than α_{k^*} is at most 2. Thus, even if all low-weight true pairs are provided as seeds, no fake pair will be matched by the PGM algorithm with threshold $r = 3$. This intuition is made precise by the following lemma.

Lemma 8. *Denote $\overline{P}_{\ge k^*} = \{u : w_u \in [0, (1+2\delta)\alpha_{k^*-1}]\}$. Fix any two distinct vertices $u, v \in \overline{P}_{\ge k^*+1}$. Then for all sufficiently large n,*

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{\ge k^*}\right| \le 2\right\} \ge 1 - n^{-2}. \tag{23}$$

Although the PGM algorithm may fail to match some true pairs with very few common 1-hop neighbors, it is expected to match the true pair with at least three 1-hop witnesses. In particular, let us recursively define

$$S_0 = P_{k^*}, \quad S_h = \{u : u \in P_{h+k^*},\ |\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap S_{h-1}| \ge 3\} \quad \text{for } h \ge 1.$$

Note that $S_0 = P_{k^*}$ has been correctly matched based on Lemma 7 in the previous step. Also, once the true pairs in S_{h-1} are added into the set of matched pairs, the PGM algorithm with threshold $r = 3$ can use the vertices in S_{h-1} as new seeds to match vertices in S_h correctly. Therefore, all the true pairs in S_h for any $h \ge 1$ can be correctly matched. Thus, to show the PGM matches many true pairs, it suffices to bound from below the size of S_h for $h \le h^*$, which is done by the following theorem.

Lemma 9. *Let* $\widetilde{w} \triangleq \left(\frac{192\overline{w}\ln 2}{Cs^2}\right)^{1/(3-\beta)}$. *Define* h^* *such that* $\widetilde{w} \le \alpha_{k^*+h^*} < 2\widetilde{w}$. *Then for any* $1 \le h \le h^*$, *and all sufficiently large* n,

$$\mathbb{P}\left\{|S_h| \ge \frac{1}{2}n_{k^*+h}\right\} \ge 1 - n^{-3+o(1)}. \tag{24}$$

The proof of Lemma 9 follows by induction. Assume (24) holds for $h-1$. Then analogous to the intuition of (9), for any u in P_{k^*+h}, $\mathbb{E}\left[\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap S_{h-1}\right|\right] \approx \frac{\alpha_{k^*+h}^{3-\beta} Cs^2}{\overline{w}} \ge 4\ln 2$. Hence, we can show that $\mathbb{P}\{u \in S_h\} \ge \frac{3}{4}$, which further implies (24) holds for h by concentration.

By Lemma 9, the PGM matches at least half of true pairs in $P_{k^*+h^*}$. Note that the number of vertices in $P_{k^*+h^*}$ satisfies $n_{k^*+h^*} = Cn(\alpha_{k^*+h^*-1})^{1-\beta} \ge Cn(\widetilde{w})^{1-\beta} = \Theta(n)$, as $\widetilde{w} = \Theta(1)$. Thus, the set of matched pairs by the PGM contains a constant fraction of true pairs. More precisely, we have the following lemma.

Lemma 10. *Under the conditions of Theorem 2, for all sufficiently large* n, *with probability at least* $1 - n^{-1+o(1)}$, *the set of matched pairs in Step 10 of Algorithm 1, denoted by* $\mathcal{R}_{k^*+1}$, *contains all true pairs in* S_h *and no fake pairs in* $\widehat{Q}_{k^*+h}$ *for all* $h \ge 1$. *In particular, we have* $|\mathcal{R}_{k^*+1}| = \Theta(n)$ *with probability at least* $1 - n^{-1+o(1)}$.

7.5 Match Pairs in $\widehat{Q}_0$ using 1-hop Algorithm

Given that a large constant fraction of true pairs with weights smaller than α_0 are matched error-free, we show that all the true pairs in Q_0 are matched error-free by the 1-hop algorithm.

When we match vertices in $\widehat{Q}_0$, we use $\widehat{\mathcal{R}}$, the set of pairs matched in Step $5-10$ of Algorithm 1, as seeds. Note that all true pairs in Q_{k^*} have been proved to be matched correctly with high probability. The number of true pairs in Q_{k^*} is $\Theta(n\alpha_{k^*-1}^{1-\beta})$ and the vertex in P_0 has weight larger than n^γ. Moreover, a vertex in P_0 connects to a vertex in P_{k^*} with probability at least $\frac{\alpha_0\alpha_{k^*}}{n\overline{w}}$. Therefore, for a true pair in Q_0, to bound from below its number of 1-hop witnesses in $\widehat{\mathcal{R}}$, it suffices to consider its number of 1-hop common neighbors in P_{k^*}, which is about $n\alpha_{k^*-1}^{1-\beta} \times \frac{\alpha_0\alpha_{k^*}}{n\overline{w}} = \Theta(\alpha_{k^*}^{2-\beta}n^\gamma)$. More precisely, we have the following theorem.

Lemma 11. *Fix any vertex* $u \in P_0$. *For all sufficiently large* n,

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k^*}\right| \ge \frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{8\overline{w}}\right\} \ge 1 - n^{-4}. \tag{25}$$

We caution the reader that even though the true pair (u,u) may have more 1-hop witnesses in Q_{k^*+1} than Q_{k^*}, we cannot consider its number of 1-hop common neighbors in P_{k^*+1}, because the PGM algorithm only matches a subset of the true pairs in Q_{k^*+1} and this subset is random and may incur dependency issues to the analysis.

Next we study fake pairs. Note that with high probability $\widehat{\mathcal{R}}$ contains no fake pair in $\bigcup_{k\ge1}\widehat{Q}_k$. Therefore, on the event that $\widehat{P}_k \subset \overline{P}_k$ for all $k \ge 1$, all the matched pairs in $\widehat{\mathcal{R}}$ is contained by the set of true pairs in $\overline{R} \times \overline{R}$, where $\overline{R} = \bigcup_{k\ge1}\overline{P}_k = \{i : w_i \in [0, (1+2\delta)n^\gamma]\}$. Therefore, for a fake pair in $\widehat{Q}_0$, to bound from above its the number of 1-hop witnesses in $\widehat{\mathcal{R}}$, it suffices to bound its number of 1-hop common neighbors in $\overline{R}$, which is done in the following lemma. Again, to resolve the potential dependency between $\widehat{Q}_0$ and graphs G_1, G_2, we state the lemma for a fake pair in $\overline{Q}_0$, which contains $\widehat{Q}_0$ with high probability.

Lemma 12. *Denote* $\overline{R} = \{i : w_i \in [0, (1+2\delta)n^\gamma]\}$. *Fix any two distinct vertices* $u, v \in \overline{P}_0$. *For all sufficiently large* n,

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{R}\right| \le 4\kappa n^{\gamma(3-\beta)} s^2\right\} \ge 1 - n^{-4}, \tag{26}$$

where $\kappa = \frac{(1+2\delta)^2 2^{5-\beta} C}{(2^{3-\beta}-1)\overline{w}}$.

To see how (26) follows, note that a vertex in P_k becomes a common 1-hop neighbor of the fake pair (u, v) with probability at most $\left(\frac{\alpha_k w_{\max}}{n\overline{w}}\right)^2 \le \frac{\alpha_k^2}{n\overline{w}}$. Since there are $\Theta(n\alpha_k^{1-\beta})$ true pairs in Q_k, the number of common 1-hop neighbors in $\overline{R}$ is on the order of $\sum_{k=1}^{K} \frac{\alpha_k^{3-\beta}}{\overline{w}} = \Theta\left(n^{\gamma(3-\beta)}\right)$.

Recall that $\overline{P}_0 \subset P_0 \cup P_1$. Thus for any fake pair $(u, v) \in \overline{Q}_0$, the two corresponding true pairs $(u, u), (v, v) \in Q_0 \cup Q_1$. If one of them is in Q_1, then it has already been matched in $\widehat{Q}_1$ by Lemma 4. If one of them is in Q_0, since $\alpha_{k^*}^{2-\beta} n^\gamma = \Theta\left(n^\gamma (\log n)^{(2-\beta)/(3-\beta)}\right) \gg n^{\gamma(3-\beta)}$ in view of $2 < \beta < 3$, it has more 1-hop witnesses than the fake pair (u, v). Thus, we expect that the 1-hop algorithm can match all the true pairs in $\widehat{Q}_0$ error-free. More precisely, we have the following lemma.

Lemma 13. *Under the conditions of Theorem 2, for all sufficiently large* n, *with probability at least* $1 - n^{-2.5}$, *the set of matched pairs in Step 11 of Algorithm 1, denoted by* $\mathcal{R}_0$, *contains all true pairs in* Q_0 *and no fake pairs in* $\widehat{Q}_0$.

7.6 Proof of Theorem 2

Due to Lemma 10 and $\mathcal{R}_{k^*+1} \subset \mathcal{R}$, the set of matched pairs by Algorithm 1 contains $\Theta(n)$ true pairs with probability at least $1 - n^{-1+o(1)}$. Combining Lemma 4, Lemma 7, Lemma 10 and Lemma 13, $\mathcal{R}$ contains no fake pairs with probability at least $1 - n^{-1+o(1)}$.

8 CONCLUSION

In this paper, we propose an efficient seeded algorithm for matching graphs with power-law degree distributions. Theoretically, under the Chung-Lu model with power-law exponent $2 < \beta < 3$ and max degree $\Theta(\sqrt{n})$, we show that as soon as $D > \frac{4-\beta}{3-\beta}$, by optimally choosing the first slice, our algorithm correctly matches a constant fraction of true pairs without any error with high probability, provided with only $\Omega((\log n)^{4-\beta})$ initial seeds. This result achieves an exponential reduction in the seed size requirement, as the previously best known result requires $n^{1/2+\epsilon}$ initial seeds. Empirically, numerical experiments in both synthetic and real power-law graphs further demonstrate that our algorithm significantly outperforms the state-of-the-art algorithms. These results uncover the enormous power of D-hops in seeded graph matching under power-law graphs.

Our work can be extended along several future directions. First, our work focuses on the challenging scenario when all seeds are uniformly chosen. We expect that it would be easier to match graphs when more high-degree vertex-pairs are chosen as seeds. It would be interesting to extend our algorithm to such cases with non-uniform seeds and reduce the existing seed requirement $\Omega(n^\epsilon)$ in [5, 7]. Second, the Chung-Lu model may not well capture some properties of real networks, such as the clustering coefficients and the abundance of triangles [28]. An alternative power-law model is the celebrated preferential attachment model [2]. It remains open whether similar performance guarantees for D-hop algorithm can be shown under the preferential attachment model to reduce the existing seed requirement $\Omega(n/\log(n))$ in [19]. Finally, another interesting and important future direction to further explore the power of D-hops in matching power-law graphs without seeds.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their suggestions and comments. L. Yu and J. Xu are supported by the NSF Grant IIS-1932630.

REFERENCES

[1] David Avis. 1983. A Survey of Heuristics for the Weighted Matching Problem. *Networks* 13, 4 (1983), 475–493.
[2] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512.
[3] Albert-László Barabási. 2016. *Network Science*. Cambridge University Press, Cambridge. http://barabasi.com/networksciencebook/
[4] Alexander C Berg, Tamara L Berg, and Jitendra Malik. 2005. Shape Matching and Object Recognition Using Low Distortion Correspondences. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE Computer Society, 26–33.
[5] Karl Bringmann, Tobias Friedrich, and Anton Krohmer. 2014. De-anonymization of Heterogeneous Random Graphs in Quasilinear Time. In *European Symposium on Algorithms*. Springer, 197–208.
[6] T. Caelli and S. Kosinov. 2004. An Eigenspace Projection Clustering Method for Inexact Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 4 (2004), 515–519.
[7] Carla-Fabiana Chiasserini, Michele Garetto, and Emilio Leonardi. 2016. Social Network De-Anonymization Under Scale-Free User Relations. *IEEE/ACM Transactions on Networking* 24, 6 (2016), 3756–3769.
[8] Fan Chung and Linyuan Lu. 2004. The Average Distance in a Random Graph with Given Expected Degrees. *Internet Mathematics* 1, 1 (2004), 91–113.
[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law Distributions in Empirical Data. *SIAM review* 51, 4 (2009), 661–703.
[10] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. 2004. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence* 18, 03 (2004), 265–298.
[11] Timothee Cour, Praveen Srinivasan, and Jianbo Shi. 2006. Balanced Graph Matching. In *Advances in Neural Information Processing Systems 19*. 313–320.
[12] Devdatt P. Dubhashi and Alessandro Panconesi. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511581274
[13] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. 2020. Spectral Graph Matching and Regularized Quadratic Relaxations: Algorithm and Theory. In *International Conference on Machine Learning*. PMLR, 2985–2995.
[14] Marcelo Fiori, Pablo Sprechmann, Joshua Vogelstein, Pablo Muse, and Guillermo Sapiro. 2013. Robust Multimodal Graph Matching: Sparse Coding Meets Graph Matching. In *Advances in Neural Information Processing Systems 26*. 127–135.
[15] Donniell E. Fishkind, Sancar Adali, Heather G. Patsolic, Lingyao Meng, Digvijay Singh, Vince Lyzinski, and Carey E. Priebe. 2018. Seeded Graph Matching. arXiv:1209.0367 [stat.ML]
[16] Aria Haghighi, Andrew Y Ng, and Christopher D Manning. 2005. Robust Textual Inference via Graph Matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 387–394.
[17] Ehsan Kazemi, Hamed Hassani, Matthias Grossglauser, and Hassan Pezeshgi Modarres. 2016. PROPER: Global Protein Interaction Network Alignment Through Percolation Matching. *BMC bioinformatics* 17, 1 (2016), 527.
[18] Ehsan Kazemi, S Hamed Hassani, and Matthias Grossglauser. 2015. Growing a Graph Matching from a Handful of Seeds. *Proceedings of the VLDB Endowment* 8, 10 (2015), 1010–1021.
[19] Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment* 7, 5 (2014), 377–388.
[20] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[21] Joseph Lubars and R Srikant. 2018. Correcting the Output of Approximate Graph Matching Algorithms. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1745–1753.
[22] Vince Lyzinski, Donniell E. Fishkind, and Carey E. Priebe. 2013. Seeded Graph Matching for Correlated Erdős-Rényi Graphs. *Journal of Machine Learning Research* 15 (2013).
[23] Elchanan Mossel and Jiaming Xu. 2019. Seeded Graph Matching via Large Neighborhood Statistics. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1005–1014.
[24] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy*. IEEE, 111–125.

[25] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 173–187.
[26] Mark EJ Newman. 2003. The Structure and Function of Complex Networks. *SIAM review* 45, 2 (2003), 167–256.
[27] Pedram Pedarsani and Matthias Grossglauser. 2011. On the Privacy of Anonymized Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1235–1243.
[28] Ali Pinar, C. Seshadhri, and Tamara G. Kolda. 2011. The Similarity between Stochastic Kronecker and Chung-Lu Graph Models. arXiv:1110.4925 [cs.SI]
[29] Christian Schellewald and Christoph Schnörr. 2005. Probabilistic Subgraph Matching Based on Convex Relaxation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 171–186.
[30] Farhad Shirani, Siddharth Garg, and Elza Erkip. 2017. Seeded Graph Matching: Efficient Algorithms and Theoretical Guarantees. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 253–257.
[31] Rohit Singh, Jinbo Xu, and Bonnie Berger. 2008. Global Alignment of Multiple Protein Interaction Networks With Application to Functional Orthology Detection. *Proceedings of the National Academy of Sciences* 105, 35 (2008), 12763–12768.
[32] Amanda L Traud, Peter J Mucha, and Mason A Porter. 2012. Social Structure of Facebook Networks. *Physica A: Statistical Mechanics and its Applications* 16, 391 (2012), 4165–4180.
[33] Lyudmila Yartseva and Matthias Grossglauser. 2013. On the Performance of Percolation Graph Matching. In *Proceedings of the 1st ACM Conference on Online Social Networks*. ACM, 119–130.
[34] Liren Yu, Jiaming Xu, and Xiaojun Lin. 2021. Graph Matching with Partially-Correct Seeds. arXiv:2004.03816 [cs.DS]

A ALGORITHM

In this section, we present the GMWM algorithm [1] and the PGM algorithm [33] used in our PLD algorithm.

A.1 The GMWM Algorithm

After counting the number of witnesses, we can form a weighted bipartite graph G_m, with the vertex set being a collection of all vertices in G_1 and G_2, the edges connecting every possible vertex-pairs, and the weight of an edge being defined as the number of witnesses. Then, we use the GMWM algorithm shown in Algorithm 2 to find the matches in G_m with large weights. GMWM first chooses the vertex-pair with the largest weight from all candidate vertex-pairs in G_m. Then, it removes all edges adjacent to the chosen vertex-pair, and then chooses the vertex-pair with the largest weight among the remaining candidate vertex-pairs, and so on.

Algorithm 2 The Greedy Maximum Weight Matching (GMWM) Algorithm.

1: **Input:** Graph G_m, the set of matched pairs $M = \emptyset$.
2: **while** G_m contains edges **do**
3: Choose the pair (i, j) with largest weight.
4: Add (i, j) into M.
5: delete i, j and their adjacent edges from G_m.
6: **end while**
7: **Output:** The set of matched pairs M

A.2 The PGM Algorithm

The PGM algorithm proposed in [33] is shown in Algorithm 3. The algorithm iteratively matches pairs with at least r witnesses. The algorithm maintains a set M of matched pairs (which is initially the seed set $\mathcal{S}$), and a set Z of used pairs (which is initially empty). At each iteration, the algorithm uses exactly one unused but already matched pair $(i, j) \in M \setminus Z$. This vertex-pair adds one mark (i.e., one witness) to each neighboring pair, i.e., to every pair in $\Gamma_1^{G_1}(i) \times \Gamma_1^{G_2}(j)$. This vertex-pair (i, j) then becomes a used pair, which is added to Z. As soon as any vertex-pair gets at least r marks,

it is added to the set M of matched pairs. The process iterates until there are no more unused pairs in $M \setminus Z$.

Algorithm 3 The Percolation Graph Matching (PGM) Algorithm.

Input: Graphs G_1, G_2, initial seed set $\mathcal{S}$, used seed set $Z = \emptyset$, threshold r.
Let $M = \mathcal{S}$
for all vertex-pairs $(i, j) \in M \setminus Z$ **do**
  Add one mark to all candidate vertex-pairs (i', j') such that $i' \in \Gamma_1^{G_1}(i)$ and $j' \in \Gamma_1^{G_2}(j)$.
  if a vertex-pair (i', j') has at least r marks **then**
    Add (i', j') into M.
    All other candidates (i', j'') and (i'', j') are discarded from consideration.
  end if
  $Z = Z \cup \{(i, j)\}$.
end for
Output: The set of matched pairs M

B COMPUTATIONAL COMPLEXITY ANALYSIS

We analyze the computational complexity of Algorithm 1 in each step.

First, Algorithm 1 checks all the vertices degrees to construct the subgraphs $\widehat{G}_1, \widehat{G}_2$ and partition the vertices in G_1 and G_2 into slices based on vertex degrees in line 2-4 and line 9. The total time complexity of this step is $O(n)$.

We then apply the D-hop algorithm in the first slice. Searching for all D-hop neighbors of a given vertex u in the first slice takes a total of $O(n)$ time steps. The number of vertices in the first slice in $\Theta(n\alpha_1^{1-\beta})$. Thus, the complexity of counting D-hop witnesses for all vertices-pairs in the first slice-pair is $O(n^3\alpha_1^{2(1-\beta)}) = O(n^{3-2\gamma(\beta-1)})$. Since we have shown that with high probability, all the fake pairs have D-hop witnesses fewer than the threshold, we only need to sort and match at most n true pairs using GMWM and hence the complexity of the GMWM step is $O(n \log n)$.

We next apply the 1-hop algorithm in the subsequent slices. We compute the number of 1-hop witnesses via neighborhood exploration. For each matched pair in Q_{k-1}, we fetch its 1-hop neighbors of size $O(\alpha_{k-1})$ in $\widehat{G}_1$ and $\widehat{G}_2$, and then increase the number of 1-hop witnesses by 1 for $O(\alpha_{k-1}^2)$ vertex-pairs. Thus, the total complexity of our algorithm to match vertices in P_k is about $n\alpha_{k-1}^{1-\beta} \times \alpha_{k-1}^2 = O(n^{1+\gamma(3-\beta)})$.

Analogously, the PGM algorithm explores the 1-hop neighbors of each matched pair. There are at most n matched pair, and for each mathced pair, we increase the number of 1-hop witnesses by 1 for $O(\triangle^2)$ vertex-pairs, where $\triangle$ is the largest degree among G_1' and G_2'. By the definition, $\triangle$ is $O((\log n)^{\frac{1}{3-\beta}})$. Therefore, the total complexity in line 10 is $O(n(\log n)^{\frac{2}{3-\beta}})$.

Finally, there are at most n true pairs to serve as 1-hop witnesses for vertex-pairs in $\widehat{Q}_0$. For any true pair (i, i), the complexity of neighborhood exploration is $O(|\Gamma_1^{G_1}(i)||\Gamma_1^{G_2}(i)|)$. Thus, the complexity of line 11 is $\sum_{i=1}^n |\Gamma_1^{G_1}(i)||\Gamma_1^{G_2}(i)| = O(\sum_{i=1}^n w_i^2) = O(n^{1+(3-\beta)/2})$ as shown in [8, page 98].

In conclusion, by summing up the complexity for each step, the total computational complexity of our algorithm is $O\left((n^{3-2\gamma(\beta-1)} + n\log n + n^{1+\gamma(3-\beta)} + n(\log n)^{\frac{2}{3-\beta}} + n^{1+(3-\beta)/2}\right) = O\left(n^{3-2\gamma(\beta-1)}\right)$ due to $\gamma \le 1/2$ and $2 < \beta < 3$.

C PROOF

C.1 Supporting Theorems

THEOREM 3. *Chernoff Bound ([12]): Let $X = \sum_{i\in[n]} X_i$, where X_i, $i \in [n]$, are independent random variables taking values in $\{0, 1\}$. Then, for $\eta \in [0, 1]$,*

$$\mathbb{P}\{X \le (1-\eta)\mathbb{E}[X]\} \le \exp\left(-\frac{\eta^2}{2}\mathbb{E}[X]\right),\ \mathbb{P}\{X \ge (1+\eta)\mathbb{E}[X]\} \le \exp\left(-\frac{\eta^2}{3}\mathbb{E}[X]\right).$$

THEOREM 4. *Bernstein's Inequality ([12]): Let $X = \sum_{i\in[n]} X_i$, where X_i, $i \in [n]$, are independent random variables such that $|X_i| \le K$ almost surely. Then, for $t > 0$, we have*

$$\mathbb{P}\{X \ge \mathbb{E}[X] + t\} \le \exp\left(-\frac{t^2}{2(\sigma^2 + Kt/3)}\right),$$

where $\sigma^2 = \sum_{i\in[n]} \mathrm{var}(X_i)$ is the variance of X. It follows then for $\rho > 0$, we have

$$\mathbb{P}\left\{X \ge \mathbb{E}[X] + \sqrt{2\sigma^2\rho} + \frac{2K\rho}{3}\right\} \le \exp(-\rho).$$

The obtained estimate holds for $\mathbb{P}\left\{X \le \mathbb{E}[X] - \sqrt{2\sigma^2\rho} - \frac{2K\rho}{3}\right\}$ too (by considering $-X$), i.e.,

$$\mathbb{P}\left\{X \le \mathbb{E}[X] - \sqrt{2\sigma^2\rho} - \frac{2K\rho}{3}\right\} \le \exp(-\rho).$$

THEOREM 5. *([34, Theorem 6]) For $r \ge 0$, every real number $x \in (0, 1)$ and $rx \le 1$, it holds that*

$$r\log(1-x) \le \log\left(1 - \frac{rx}{2}\right).$$

THEOREM 6. *([34, Corollary 1]) Let X denote a random variable such that $X \sim \mathrm{Binom}(n, p)$. If $n \in [n_{\min}, n_{\max}]$, then for $\lambda > 0$,*

$$\mathbb{P}\left\{X \ge 2n_{\max}\alpha + \frac{4\gamma}{3}\right\} \le \exp(-\gamma) \tag{27}$$

C.2 Proof of the Main Result

First, we define some notations related to graph slicing. We count the number of vertices in the slice P_k and $\overline{P}_k$. The vertices in P_k satisfies

$$\alpha_k \le w_i \le \alpha_{k-1} \Longleftrightarrow \frac{n}{\left(\frac{(\beta-1)n^\gamma}{(\beta-2)\overline{w}2^{k-1}}\right)^{\beta-1}} - i_0 \le i \le \frac{n}{\left(\frac{(\beta-1)n^\gamma}{(\beta-2)\overline{w}2^{k}}\right)^{\beta-1}} - i_0.$$

According to the index range of the vertices, we define n_k to be the difference between the two bounds. To be more precise,

$$n_k \triangleq Cn\alpha_{k-1}^{1-\beta}, \tag{28}$$

where C throughout this paper denotes $(2^{\beta-1} - 1)\left(\frac{(\beta-2)\overline{w}}{(\beta-1)}\right)^{\beta-1}$. Moreover, we have that

$$n_k \le |P_k| \le n_k + 1 \le \frac{11}{10}n_k. \tag{29}$$

Similarly, the vertices in $\overline{P}_k$ satisfies

$$(1-2\delta)\alpha_k \le w_i \le (1+2\delta)\alpha_{k-1}.$$

Thus,

$$\begin{aligned}\left|\overline{P}_k\right| &\le \left(2^{\beta-1}(1+2\delta)^{\beta-1} - (1-2\delta)^{\beta-1}\right)\frac{n_k}{2^{\beta-1}-1} + 1\\ &\overset{(a)}{\le} \frac{\left(\frac{5}{2}\right)^{\beta-1} - \left(\frac{3}{4}\right)^{\beta-1}}{2^{\beta-1}-1} n_k + 1 \le 2n_k,\end{aligned} \tag{30}$$

where (a) follows from $\delta = \frac{1}{8}$.

The number of perfect slices, denoted by K, is

$$\log_2\left(n^\gamma\right) \le K \le 1 + \log_2\left(n^\gamma\right).$$

C.2.1 Proof of Lemma 1. First, we prove $P_k \subset \widehat{P}_k^{G_j}$ with high probability for $0 \le k \le k^*$ and $j = 1, 2$. Fix any vertex u in P_k. It suffices to show with high probability $u \in \widehat{P}_k^{G_j}$. Note that any vertex v connects to u in G_j independently with probability $p_{uv}s$, where $j = 1, 2$ and $p_{uv} = \frac{w_u w_v}{n\overline{w}}$. Thus

$$\mathbb{E}\left[\left|\Gamma_1^{G_j}(u)\right|\right] = \sum_{v \in G_j} p_{uv}s = w_u s.$$

Note that $\alpha_k \le w_u \le \alpha_{k-1}$ and $\alpha_k \ge \alpha_{k^*} \ge \left(\frac{85\overline{w}\log n}{Cs^2}\right)^{\frac{1}{3-\beta}} \ge \frac{20\log n}{\delta^2 s}$ for the choice of k^* in (4) and sufficiently large n, in view of $2 < \beta < 3$. Then, applying the Chernoff Bound in Theorem 3 with $\eta = \delta$ yields

$$\mathbb{P}\left\{\left|\Gamma_1^{G_j}(u)\right| \ge (1+\delta)\alpha_{k-1}s\right\} \le \exp\left(-\delta^2\frac{\alpha_{k-1}s}{3}\right) \le n^{-5},$$

and

$$\mathbb{P}\left\{\left|\Gamma_1^{G_j}(u)\right| \le (1-\delta)\alpha_k s\right\} \le \exp\left(-\delta^2\frac{\alpha_k s}{2}\right) \le n^{-5}.$$

Combining the last two displayed equation yields that

$$\mathbb{P}\left\{u \notin \widehat{P}_k^{G_j}\right\} \le 2n^{-5}.$$

Taking an union bound over u gives

$$\mathbb{P}\left\{P_k \subset \widehat{P}_k^{G_j}\right\} \ge 1 - \sum_{u \in P_k} \mathbb{P}\left\{u \notin \widehat{P}_k^{G_j}\right\} \ge 1 - n^{-4+o(1)}. \tag{31}$$

Next we show that $P_{\ge k^*} \subset \widehat{P}_{\ge k^*}^{G_j}$ with high probability. Fix any vertex $u \in P_k$ with $k \ge k^*$. Take a vertex $v \in P_{k^*}$ with $w_v = \alpha_{k^*-1}$. Since $w_u \le w_v$, we have $\left|\Gamma_1^{G_j}(u)\right| \overset{s.t.}{\le} \left|\Gamma_1^{G_j}(v)\right|$. Therefore,

$$\mathbb{P}\left\{u \notin \widehat{P}_{\ge k^*}^{G_j}\right\} = \mathbb{P}\left\{\left|\Gamma_1^{G_j}(u)\right| \ge (1+\delta)\alpha_{k^*-1}s\right\} \le \mathbb{P}\left\{\left|\Gamma_1^{G_j}(v)\right| \ge (1+\delta)\alpha_{k^*-1}s\right\} \le n^{-5},$$

Taking a union bound over u gives

$$\mathbb{P}\left\{P_{\ge k^*} \subset \widehat{P}_{\ge k^*}^{G_j}\right\} \ge 1 - n^{-4+o(1)}. \tag{32}$$

Second, we prove that for $0 \le k \le k^*$, with high probability $\widehat{P}_k \subset \overline{P}_k$, or equivalently, $[n]\backslash\overline{P}_k \subset [n]\backslash\widehat{P}_k$, Fix any vertex u with $w_u > (1+2\delta)\alpha_{k-1}$, applying the Chernoff Bound with $\eta = \frac{\delta}{1+2\delta}$ yields

$$\mathbb{P}\left\{\left|\Gamma_1^{G_j}(u)\right| \le (1+\delta)\alpha_{k-1}s\right\} \le \exp\left(-\delta^2\frac{\alpha_{k-1}s}{2(1+2\delta)}\right) \le n^{-5}. \tag{33}$$

For any vertex u with $w_u < (1-2\delta)\alpha_k$, applying the Chernoff Bound with $\eta = \frac{\delta}{1-2\delta}$ yields

$$\mathbb{P}\left\{\left|\Gamma_1^{G_j}(u)\right| \geq (1-\delta)\alpha_k s\right\} \leq \exp\left(-\delta^2 \frac{\alpha_k s}{3(1-2\delta)}\right) \leq n^{-5}. \tag{34}$$

Thus, we have

$$\mathbb{P}\left\{\widehat{P}_k^{G_j} \subset \overline{P}_k\right\} = \mathbb{P}\left\{[n]\backslash\overline{P}_k \subset [n]\backslash\widehat{P}_k^{G_j}\right\} \geq 1 - \sum_{u\notin\overline{P}_k} \mathbb{P}\left\{u \in \widehat{P}_k^{G_j}\right\} \geq 1 - n^{-4}, \tag{35}$$

where the last inequality holds by combining (33) and (34). Moreover,

$$\mathbb{P}\left\{\widehat{P}_{\geq k^*}^{G_j} \subset \overline{P}_{\geq k^*}\right\} = \mathbb{P}\left\{[n]\backslash\overline{P}_{\geq k^*} \subset [n]\backslash\widehat{P}_{\geq k^*}^{G_j}\right\} \geq 1 - \sum_{u:w_u > (1+2\delta)\alpha_{k^*-1}} \mathbb{P}\left\{u \in \widehat{P}_{\geq k^*}^{G_j}\right\} \geq 1 - n^{-4}, \tag{36}$$

where the last inequality holds by (33).

Then, combining (31) and (35) with the union bound yields that $\mathbb{P}\left\{Q_k \subset \widehat{Q}_k \subset \overline{Q}_k\right\} \geq 1 - n^{-4+o(1)}$ for $0 \leq k \leq k^*$. Similarly, combining (32) and (36) with a union bound yields that $\mathbb{P}\left\{Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}\right\} \geq 1 - n^{-4+o(1)}$.

Finally, since $\underline{V} = \bigcup_{k\geq 1} P_k$, $V = \bigcup_{k\geq 1} \widehat{P}_k^{G_j}$ and $\overline{V} = \bigcup_{k\geq 1} \overline{P}_k$, combining (31), (32), (35), and (36) with the union bound, we have

$$\mathbb{P}\left\{\underline{G}_j \subset \widehat{G}_j \subset \overline{G}_j\right\} = \mathbb{P}\left\{\underline{V} \subset V_j \subset \overline{V}\right\} \geq 1 - n^{-3+o(1)}.$$

C.2.2 Proof of Lemma 2. Note that $\underline{G}_1 \wedge \underline{G}_2$, $\overline{G}_1$, and $\overline{G}_2$ are graphs that are edge-sampled from G_0 with probability s^2, s, s, respectively. Thus, we let G denote a graph obtained by sampling each edge of G_0 independently with probability $t = \Theta(1)$ and $\overline{G}$ denote a subgraph of G induced by the vertex set $\overline{V} = \{u : w_u \in [0, (1+2\delta)n^\gamma]\}$. Fix a vertex $u \in P_1$, we first study its number of d-hop neighbors in each slice in $\overline{G}$. Then, we can arrive at Lemma 2 by selecting the corresponding parameters. To be more precise, we define $\Gamma_{d,k}^{\overline{G}}(u) = \Gamma_d^{\overline{G}}(u) \cap \overline{P}_k$ and $N_{d,k}^{\overline{G}}(u) = \bigcup_{1\leq j\leq d} \Gamma_{j,k}^{\overline{G}}(u)$. We bound $\Gamma_{d,k}^{\overline{G}}(u)$ and $N_{d,k}^{\overline{G}}(u)$ by the following lemma.

Lemma 14. *Fix any vertex $u \in P_1$, and let Ω_d denote the event such that the followings hold simultaneously for $k = 1, \ldots, K$:*

$$\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \geq 2^{(k-1)(\beta-2)} \left(\frac{(1-2\delta)^2 C \cdot t}{12 \cdot \overline{w}}\right)^d n^{\gamma(3-\beta)d} \triangleq \Gamma_{\min}(d,k), \tag{37}$$

$$\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \leq 2^{(k-1)(\beta-2)} \kappa^d n^{\gamma(3-\beta)d} \triangleq \Gamma_{\max}(d,k), \tag{38}$$

$$\left|N_{d,k}^{\overline{G}}(u)\right| \leq 2^{(k-1)(\beta-2)+1} \kappa^d n^{\gamma(3-\beta)d}, \tag{39}$$

where $\kappa = \frac{(1+2\delta)^2 2^{5-\beta} C}{(2^{3-\beta}-1)\overline{w}}$. Suppose γ and D are chosen such that condition (10) holds. Then, for all $1 \leq d \leq D$ and sufficiently large n,

$$\mathbb{P}\{\Omega_d\} \geq 1 - (4^d - 1)n^{-4}. \tag{40}$$

Remark 2. The intuition behind Lemma 14 is as follows. Recall that q_d, the probability that a vertex of $\Theta(1)$ weight lies in the d-hop neighborhood of a vertex in the first slice, is on the order of $n^{\gamma[(3-\beta)(d-1)+1]-1}$ in view of (7). Note that the weight of vertices in P_k is about α_k, and the size of P_k is $\Theta(n\alpha^{1-\beta})$. Thus, the expected number of vertices in P_k that are d-hop neighbors of a given

vertex in the first slice is roughly $nq_d\alpha_k^{2-\beta} \approx 2^{(k-1)(\beta-2)}n^{\gamma(3-\beta)d}$. Hence, we expect (37)– (39) to hold with high probability by concentration.

Before proving Lemma 14, we first show how to apply Lemma 14 to prove Lemma 2. By setting $\delta = 0$ and $t = s^2$, we have $\overline{G} = \underline{G}_1 \wedge \overline{G}_2$. Thus, (37) with $k = \lceil \log_2(n^\gamma) \rceil$ and $d = D$ leads to the desired conclusion (16). Moreover, there are at most c slices in $\{i : w_i \le c\}$. By setting $\delta = \frac{1}{8}$, $d = D-1$, $\overline{G} = \overline{G}_j$ (i.e., $t = s$), (39) with $\log_2(n^\gamma/c) \le k \le K \le \log_2(n^\gamma) + 1$, we have

$$\sum_{k=\lfloor \log_2(n^\gamma/c)\rfloor}^{K} 2^{(k-1)(\beta-2)+1}\kappa^{D-1}n^{\gamma(3-\beta)(D-1)} \le 2c\kappa^{D-1}n^{\gamma((3-\beta)(D-2)+1)} = N_{\max},$$

where $N_{\max}$ is given in (17). Thus, we prove the desired conclusion (17).

We then present the proof of Lemma 14.

proof of Lemma 14. Fix a vertex u in P_1, we study its d-hop neighborhood in $\overline{G}$ from $d = 1$.

For $d = 1$: For each vertex $i \in \overline{P}_k$, define an indicator variable

$$x_i^k = \mathbf{1}_{\left\{i \in \Gamma_1^{\overline{G}}(u)\right\}}.$$

In other words, $x_i^k = 1$ if i is connected to u in $\overline{G}$, and $x_i^k = 0$ otherwise. Since $u \in P_1$, it follows that

$$p_{\min}^k = (1-2\delta)\frac{\alpha_k\alpha_1}{n\overline{w}}t \le \mathbb{P}\left\{x_i^k = 1\right\} \le (1+2\delta)\frac{\alpha_{k-1}\alpha_0}{n\overline{w}}t = p_{\max}^k.$$

Then, we have $\left|\Gamma_{1,k}^{\overline{G}}(u)\right| = \sum_{i\in\overline{P}_k} x_i^k$ and x_i^k's are independent. Recall that $n_k = Cn\alpha_{k-1}^{1-\beta}$ in view of (28) and $n_k \le \left|\overline{P}_k\right| \le 2n_k$ in view of (30). Thus

$$n_k p_{\min}^k = (1-2\delta)C\frac{\alpha_{k-1}^{2-\beta}\alpha_1}{2\overline{w}}t = (1-2\delta)C\frac{n^{\gamma(3-\beta)}}{4\cdot 2^{(k-1)(2-\beta)}\overline{w}}t,$$
$$n_k p_{\max}^k = (1+2\delta)C\frac{\alpha_{k-1}^{2-\beta}\alpha_0}{\overline{w}}t = (1+2\delta)C\frac{n^{\gamma(3-\beta)}}{2^{(k-1)(2-\beta)}\overline{w}}t.$$

Hence, applying Chernoff Bound in Theorem 3 with $\eta = \frac{1}{2}$ yields that

$$\mathbb{P}\left\{\left|\Gamma_{1,k}^{\overline{G}}(u)\right| \le (1-2\delta)\frac{Cn^{\gamma(3-\beta)}t}{8\cdot 2^{(k-1)(2-\beta)}\overline{w}}\right\} \le \mathbb{P}\left\{\mathrm{Binom}\left(n_k, p_{\min}^k\right) \le \frac{1}{2}n_k p_{\min}^k\right\} \overset{(a)}{\le} n^{-4},$$

$$\mathbb{P}\left\{\left|\Gamma_{1,k}^{\overline{G}}(u)\right| \ge (1+2\delta)\frac{3Cn^{\gamma(3-\beta)}t}{2^{(k-1)(2-\beta)}\overline{w}}\right\} \le \mathbb{P}\left\{\mathrm{Binom}\left(2n_k, p_{\max}^k\right) \le 3n_k p_{\max}^k\right\} \overset{(b)}{\le} n^{-4},$$

where (a) and (b) hold because $n_k p_{\max}^k \ge n_k p_{\min}^k \ge (1-2\delta)\frac{Cn^{\gamma(3-\beta)}t}{4\cdot\overline{w}} \ge 108\log n$ for sufficiently large n.

We also have $\mathbb{P}\left\{\left|N_{1,k}^{\overline{G}}(u)\right| \ge 3n_k p_{\max}^k\right\} \le n^{-4}$ due to $N_{1,k}^{\overline{G}}(u) = \Gamma_{1,k}^{\overline{G}}(u)$. Finally, taking the union bound leads to (40) for $d = 1$.

For $2 \le d \le D$: We first count the d-hop neighbors conditional on the $(d-1)$-hop neighborhood of u such that Ω_{d-1} holds. The high-level idea is as follows. After the conditioning, every vertex i outside the $(d-1)$-hop neighborhood of u will become a d-hop neighbor by connecting to at least one of the $(d-1)$-hop neighbors v of u. These edge connections are still independently generated across different v and i according to the Chung-Lu model.

We first bound $\left|\Gamma^{\overline{G}}_{d,k}(u)\right|$ from below. For each vertex $i \in \overline{P}_k \setminus \left(N^{\overline{G}}_{d-1,k}(u)\right) \triangleq P'_k$, define an indicator variable

$$y_i^k = \mathbf{1}_{\left\{\exists v \in \Gamma^{\overline{G}}_{d-1}(u):\, i \in \Gamma^{\overline{G}}_1(v)\right\}}.$$

In other words, $y_i^k = 1$ if i is connected to at least one $(d-1)$-hop neighbor of u in $\overline{G}$, and $y_i^k = 0$ otherwise. Thus, we have $\left|\Gamma^{\overline{G}}_{d,k}(u)\right| = \sum_{i \in P'_k} y_i^k$, and y_i^k's are independent across different i conditional on Ω_{d-1}.

Note that $\Gamma^{\overline{G}}_{d-1,1}(u) \subset \Gamma^{\overline{G}}_{d-1}(u)$. Thus, we can bound $\mathbb{P}\left\{y_i^k = 1 | \Omega_{d-1}\right\}$ from below by considering the possible edge connections between i and vertices in $\Gamma^{\overline{G}}_{d-1,1}(u)$. More precisely, we get that

$$\begin{aligned}
\mathbb{P}\left\{y_i^k = 1 \mid \Omega_{d-1}\right\} \ge& \mathbb{P}\left\{\exists v \in \Gamma^{\overline{G}}_{d-1,1}(u) : i \in \Gamma^{\overline{G}}_1(v) \mid \Omega_{d-1}\right\} \\
\overset{(a)}{\ge}& 1 - (1 - p_{vi})^{\Gamma_{\min}(d-1,1)} \\
\ge& 1 - \left(1 - (1-2\delta)^2 \frac{\alpha_k \alpha_1}{n\overline{w}} t\right)^{\Gamma_{\min}(d-1,1)} \\
\overset{(b)}{\ge}& \frac{(1-2\delta)^2}{2} \Gamma_{\min}(d-1,1) \frac{\alpha_k \alpha_1 t}{n\overline{w}} \\
=& \frac{3}{2^k Cn} \left(\frac{(1-2\delta)^2 C \cdot t}{12 \cdot \overline{w}}\right)^d n^{\gamma((3-\beta)(d-1)+2)} \triangleq p^{k,d}_{\min}.
\end{aligned}$$

where (a) holds because $\left\{i \notin \Gamma^{\overline{G}}_1(v)\right\}$ are independent across v; (b) follows from Theorem 5.

Now, to bound $\left|\Gamma^{\overline{G}}_{d,k}(u)\right|$ from below, we also need a lower bound to $|P'_k|$, or equivalently an upper bound to $\left|N^{\overline{G}}_{d-1,k}(u)\right|$. Since we have conditioned on the $(d-1)$-hop neighborhood of u such that event Ω_{d-1} holds. It follows from (39) that

$$\begin{aligned}
\left|N^{\overline{G}}_{d-1,k}(u)\right| \le& 2^{(k-1)(\beta-2)+1} \kappa^{d-1} n^{\gamma(3-\beta)(d-1)} \\
=& 2\kappa^{d-1} n^{\gamma((3-\beta)(d-2)+1)} \alpha_{k-1}^{1-\beta} \\
\overset{(a)}{\le}& \frac{C}{10} n \alpha_{k-1}^{1-\beta} \le \frac{1}{9} n_k,
\end{aligned}$$

where (a) holds due to the condition (10). Thus, we have $\left|P'_k\right| \ge |P_k| - \left|N^{\overline{G}}_{d-1,k}(u)\right| \ge \frac{8}{9} n_k$.

Note that for sufficiently large n,

$$\frac{8}{9} n_k p^{k,d}_{\min} = \frac{4}{3 \cdot 2^{(k-1)(2-\beta)}} \left(\frac{(1-2\delta)^2 Ct}{12 \cdot \overline{w}}\right)^d n^{\gamma(3-\beta)d} = \frac{4}{3} \Gamma_{\min}(d,k) \ge 128 \log n.$$

Thus, we apply the Chernoff Bound in Theorem 3 with $\eta = \frac{1}{4}$ and get

$$\mathbb{P}\left\{\left|\Gamma^{\overline{G}}_{d,k}(u)\right| \le \Gamma_{\min}(d,k) \mid \Omega_{d-1}\right\} \le \mathbb{P}\left\{\mathrm{Binom}\left(\frac{8}{9} n_k, p^{k,d}_{\min}\right) \le \Gamma_{\min}(d,k) \mid \Omega_{d-1}\right\} \le n^{-4}.$$

Next, we bound $\left|\Gamma_{d,k}^{\overline{G}}(u)\right|$ from above. To this end, we bound $\mathbb{P}\left\{\overline{y}_i^k = 1|\Omega_{d-1}\right\}$ from above and get

$$\begin{aligned}\mathbb{P}\left\{\overline{y}_i^k = 1|\Omega_{d-1}\right\} &\overset{(a)}{\leq} \sum_{l=1}^{K} \mathbb{P}\left\{\exists j \in \Gamma_{d-1,l}^{\overline{G}}(u) : i \in \Gamma_1^{\overline{G}}(j) \mid \Omega_{d-1}\right\} \\ &\overset{(b)}{\leq} (1+2\delta)^2 \sum_{l=1}^{K} \Gamma_{\max}(d-1,l)\frac{\alpha_{k-1}\alpha_{l-1}}{n\overline{w}} \\ &= (1+2\delta)^2 \frac{\kappa^{d-1} n^{\gamma((3-\beta)(d-1)+2)}}{2^{k-1}n\overline{w}} \sum_{l=1}^{K} 2^{(l-1)(\beta-3)} \\ &\leq \frac{\kappa^d n^{\gamma((3-\beta)(d-1)+2)}}{2^{k+1}Cn} \triangleq p_{\max}^{k,d}, \end{aligned} \tag{41}$$

where (a) follow from the union bound; (b) holds due to the union bound and event Ω_{d-1}; (b) follows from $(1+x)^r \geq 1+rx$ for every integer $r \geq 0$ and every real number $x \geq -2$; and the last inequality follows from the definition of $\kappa = \frac{(1+2\delta)^2 2^{5-\beta} C}{(2^{3-\beta}-1)\overline{w}}$.

Also, note that $P_k' \subset \overline{P}_k$ and thus $|P_k'| \leq |\overline{P}_k| \leq 2n_k$. For sufficiently large n, we have

$$2n_k p_{\max}^{k,d} = 2^{(k-1)(\beta-2)-1}\kappa^d n^{\gamma(3-\beta)d} = \frac{1}{2}\Gamma_{\max}(d,k).$$

Hence, applying Chernoff Bound in Theorem 3 with $\eta = 1$ yields that

$$\mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \geq \Gamma_{\max}(d,k) \mid \Omega_{d-1}\right\} \leq \mathbb{P}\left\{\mathrm{Binom}\left(2n_k, p_{\max}^{k,d}\right) \geq \Gamma_{\max}(d,k)\right\} \leq n^{-4}.$$

Induction: Finally, we prove (40) by induction.
For $d = 1$, we have proved that (40) holds. Suppose that (40) holds for $d-1$. Then we have

$$\mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \leq \Gamma_{\min}(d,k)\right\} \leq \mathbb{P}\left\{\Omega_{d-1}^c\right\} + \mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \leq \Gamma_{\min}(d,k) \mid \Omega_{d-1}\right\}\mathbb{P}\left\{\Omega_{d-1}\right\} \leq 4^{d-1}\cdot n^{-4}. \tag{42}$$

Similarly, we get

$$\mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}}(u)\right| \geq \Gamma_{\max}(d,k)\right\} \leq 4^{d-1}\cdot n^{-4} \tag{43}$$

Since $\left|N_{d,k}^{\overline{G}}(u)\right| = \left|N_{d-1,k}^{\overline{G}}(u)\right| + \left|\Gamma_{d,k}^{\overline{G}}(u)\right|$, we take an union bound and have

$$\mathbb{P}\left\{\left|N_{d,k}^{\overline{G}}(u)\right| \geq 2^{(k-1)(\beta-2)+1}\kappa^d n^{\gamma(3-\beta)d}\right\} \leq (4^{d-1}-1)\cdot n^{-4} + 4^{d-1}\cdot n^{-4} = (2\cdot 4^{d-1}-1)n^{-4}. \tag{44}$$

Combining (42), (43) and (44) with an union bound, we prove that (40) holds for any $1 \leq k \leq K$ and $1 \leq d \leq D$.

□

C.2.3 Proof of Lemma 3. Note that

$$\begin{aligned} N_{D,k}^{\overline{G}_1}(u) \cap N_{D,k}^{\overline{G}_2}(v) &\subset \left(\Gamma_{D,k}^{\overline{G}_1}(u) \cup N_{D-1,k}(u,v)\right) \cap \left(\Gamma_{D,k}^{\overline{G}_2}(v) \cup N_{D-1,k}(u,v)\right) \\ &= \left(\Gamma_{D,k}^{\overline{G}_1}(u) \cap \Gamma_{D,k}^{\overline{G}_2}(v)\right) \cup N_{D-1,k}(u,v), \end{aligned} \tag{45}$$

where $N_{D-1,k}(u,v) = N^{\overline{G}_1}_{D-1,k}(u) \cup N^{\overline{G}_2}_{D-1,k}(v)$. Since we have already obtained the upper bounds of $\left|N^{G_1}_{D-1,k}(u)\right|$ and $\left|N^{G_2}_{D-1,k}(v)\right|$ by Lemma 14 by letting $\overline{G}$ to be either $\overline{G}_1$ or $\overline{G}_2$, it remains to bound from above $\left|\Gamma^{\overline{G}_1}_{D,k}(u) \cap \Gamma^{\overline{G}_2}_{D,k}(v)\right|$, which is done in the following lemma.

Lemma 15. *Suppose γ and D are chosen such that condition (10) holds. Fix any two distinct vertices $u, v \in \overline{P}_1$, for all $1 \le d \le D$, $k = 1, \ldots, K$, and sufficiently large n,*

$$\mathbb{P}\left\{\left|\Gamma^{\overline{G}_1}_{d,k}(u) \cap \Gamma^{\overline{G}_2}_{d,k}(v)\right| \le \Psi(d,k)\right\} \ge 1 - \frac{2\cdot 4^d}{3}\cdot n^{-4}, \tag{46}$$

where

$$\Psi(d,k) = \frac{\kappa^2\Gamma^2_{\max}(d-1,1)n^{\gamma(5-\beta)}}{2^{(k-1)(3-\beta)}Cn} + \frac{6\Gamma_{\max}(d-1,1)\log n}{2^{(k-1)(2-\beta)}}$$

with $\Gamma_{\max}(d-1,1) = \kappa^{d-1}n^{\gamma(3-\beta)(d-1)}$ as defined in (38) and $\kappa = \frac{(1+2\delta)^2 2^{5-\beta}C}{(2^{3-\beta}-1)\overline{w}}$.

Remark 3. We provide an intuitive explanation on the first term of $\Psi(d,k)$. Fix a vertex u. Recall that $\Gamma_{\max}(d-1,l)$ is an upper bound of its $(d-1)$-hop neighbors in P_l by Lemma 14. Thus, a vertex i in P_k connects to at least one $(d-1)$-hop neighbor of u with probability at most $\sum_{l=1}^{K}\Gamma_{\max}(d-1,l)\frac{\alpha_k\alpha_l}{n\overline{w}} \approx \kappa\Gamma_{\max}(d-1,1)\alpha_k n^{\gamma-1}$, where the approximation holds because $l=1$ is the dominating term in the summation. Moreover, there are $\Theta(n\alpha_k^{1-\beta})$ vertices in the slice P_k. Thus, for a fake pair (u,v), its number of common d-hop neighbors in P_k is about $\kappa^2\Gamma^2_{\max}(d-1,1)n^{2\gamma-1}\alpha_k^{3-\beta}$, which gives rise to the first term of $\Psi(d,k)$.

Before proving Lemma 15, we first show how to apply Lemma 15 to prove Lemma 3. combining (45), (39), and (46) yields that

$$\mathbb{P}\left\{\left|\Gamma^{\overline{G}_1}_{d,k}(u) \cap \Gamma^{\overline{G}_2}_{d,k}(v)\right| \le \Psi(d,k) + 2N_{\max}(d-1,k)\right\} > 1 - n^{-4+o(1)}.$$

Next we set $d = D$ and sum over k for all the slices P_k with weight at most $\frac{15}{s}\log n$, i.e., $\alpha_k \le \frac{15}{s}\log n$. In particular, we have $k \ge k_0 \triangleq \lfloor \log_2(\frac{n^\gamma s}{15\log n})\rfloor$ and

$$\begin{aligned}
&\sum_{k=k_0}^{K}\Psi(D,k) + 2N_{\max}(D-1,k)\\
\le&\sum_{k=k_0}^{K}\frac{\kappa^2\Gamma^2_{\max}(D-1,1)n^{\gamma(5-\beta)}}{2^{(k-1)(3-\beta)}Cn} + \frac{\Gamma_{\max}(D-1,1)}{2^{(k-1)(2-\beta)}}6\log n + \frac{4\kappa^{D-1}n^{\gamma(3-\beta)(D-1)}}{2^{(k-1)(2-\beta)}}\\
\le&\frac{2^{3-\beta}\kappa^2\Gamma^2_{\max}(D-1,1)n^{\gamma(5-\beta)}}{(2^{3-\beta}-1)2^{(k_0-1)(3-\beta)}Cn} + \frac{2^{\beta-2}}{2^{\beta-2}-1}\frac{\Gamma_{\max}(D-1,1)}{2^{(K-1)(2-\beta)}}6\log n + \frac{2^{\beta-2}}{2^{\beta-2}-1}\frac{4\kappa^{D-1}n^{\gamma(3-\beta)(D-1)}}{2^{(K-1)(2-\beta)}}\\
\le&\frac{2^{3-\beta}\kappa^{2D}n^{2\gamma((3-\beta)(D-1)+1)}}{(2^{3-\beta}-1)Cn}\left(\frac{15}{s}\log n\right)^{3-\beta} + \frac{2^{\beta-2}}{2^{\beta-2}-1}\kappa^{D-1}n^{(\gamma(3-\beta)(D-2)+1)}(4+6\log n) = \Psi_{\max},
\end{aligned}$$

where $\Psi_{\max}$ is given in (19). Thus, we prove the desired conclusion (19).

Next we present the proof of Lemma 15.

Proof of Lemma 15. Fix two distinct vertices u, v in $\overline{P}_1$, we study their common d-hop neighborhood from $d = 1$.

For $d = 1$: For each vertex $i \in \overline{P}_k$, define an indicator variable

$$x_i^k = \mathbf{1}_{\left\{i \in \Gamma_1^{\overline{G}_1}(u) \cap \Gamma_1^{\overline{G}_2}(v)\right\}}.$$

In other words, $x_i^k = 1$ if i is connected to u in $\overline{G}_1$ and v in $\overline{G}_2$, and $x_i^k = 0$ otherwise. Then, we have $\left|\Gamma_{1,k}^{\overline{G}_1}(u) \cap \Gamma_{1,k}^{\overline{G}_2}(v)\right| = \sum_{i \in \overline{P}_k} x_i^k$. Since $w_u, w_v \in [(1-2\delta)\alpha_1, (1+2\delta)\alpha_0]$, it follows that

$$\mathbb{P}\left\{x_i^k = 1\right\} \le \left((1+2\delta)^2 \frac{\alpha_{k-1}\alpha_0}{n\overline{w}}\right)^2 \triangleq p_{\max}^k.$$

Hence, we have

$$\left|\Gamma_{1,k}^{\overline{G}_1}(u) \cap \Gamma_{1,k}^{\overline{G}_2}(v)\right| \overset{s.t.}{\le} \text{Binom}\left(\left|\overline{P}_k\right|, p_{\max}^k\right).$$

Recall $n_k = Cn\alpha_{k-1}^{1-\beta}$ in view of (28) and $\left|\overline{P}_k\right| \le 2n_k$ in view of (30). Hence,

$$2n_k p_{\max}^k = (1+2\delta)^4 \frac{2C\alpha_{k-1}^{3-\beta} n^{2\gamma}}{\overline{w}^2 n}.$$

Hence, we apply Lemma 6 with $\lambda = 4\log n$, and get

$$\mathbb{P}\left\{\left|\Gamma_{1,k}^{\overline{G}_1}(u) \cap \Gamma_{1,k}^{\overline{G}_2}(v)\right| \ge \frac{4(1+2\delta)^4 C\alpha_{k-1}^{3-\beta} n^{2\gamma}}{\overline{w}^2 n} + \frac{16}{3}\log n\right\} \le n^{-4}.$$

Since $\Gamma_{\max}(0,1) = 1$, we have $\Psi(1,k) = \frac{\kappa^2 \alpha_{k-1}^{3-\beta} n^{2\gamma}}{Cn} + 6\log n$. Thus, (46) holds for $d = 1$.

For $2 \le d \le D$: We first count the d-hop neighbors conditional on the $(d-1)$-hop neighborhood of u and v. We use Ω_d^* to denote the event that Ω_{d-1} with $\overline{G} = \overline{G}_1, \overline{G}_2$ hold, and for all $k = 1, \ldots, K$,

$$\left|\Gamma_{d,k}^{\overline{G}_1}(u) \cap \Gamma_{d,k}^{\overline{G}_2}(v)\right| \le \Psi(d,k),$$

with $\Psi(d,k)$ defined in Lemma 15.

Conditioning on Ω_{d-1}^*, note that there are two possible cases under which each true pair (i,i) becomes a common d-hop neighbor of (u,v). One case is that i connects to some common $(d-1)$-hop neighbors of (u,v) in both $\overline{G}_1$ and $\overline{G}_2$. The other case is that i connects to different $(d-1)$-hop neighbors of (u,v) in $\overline{G}_1$ and $\overline{G}_2$, respectively.

For each vertex $i \in \overline{P}_k \setminus N_{D-1}(u,v)$, define two indicator variables

$$y_i^k = \mathbf{1}_{\left\{i \in \Gamma_d^{\overline{G}_1}(u), i \in \Gamma_d^{\overline{G}_2}(v)\right\}},$$
$$z_i^k = \mathbf{1}_{\left\{\exists j \in \Gamma_{d-1}^{\overline{G}_1}(u) \cap \Gamma_{d-1}^{\overline{G}_2}(v):\, i \in \Gamma_1^{\overline{G}_1}(j)\right\}}.$$

In other words, $y_i^k = 1$ if i is a d-hop neighbor of u in $\overline{G}_1$ and v in $\overline{G}_2$, and $y_i^k = 0$ otherwise. Similarly, $z_i^k = 1$ if i is connected to at least one common $(d-1)$-hop neighbor of (u,v) in both $\overline{G}_1$ and $\overline{G}_2$, and $z_i^k = 0$ otherwise. Note that $z_i^k = 1$ includes the case that i connects to some common $(d-1)$-hop neighbors of (u,v) in both $\overline{G}_1$ and $\overline{G}_2$.

We first bound $\mathbb{P}\left\{z_i^k = 1 | \Omega_{d-1}^*\right\}$ from above by

$$\mathbb{P}\left\{z_i^k = 1 \mid \Omega_{d-1}^*\right\} \overset{(a)}{\le} \sum_{l=1}^{K} \mathbb{P}\left\{\exists j \in \Gamma_{d-1}^{\overline{G}_1}(u) \cap \Gamma_{d-1}^{\overline{G}_2}(v) : i \in \Gamma_1^{\overline{G}_1}(j) \mid \Omega_{d-1}^*\right\}$$

$$\begin{aligned}
&\overset{(b)}{\leq}(1+2\delta)^2\sum_{l=1}^{K}\Psi(d-1,l)\frac{\alpha_{k-1}\alpha_{l-1}}{n\overline{w}}\\
&\leq\left(\frac{\kappa^2\Gamma_{\max}^2(d-2,1)n^{\gamma(7-\beta)}}{2^{k-1}Cn^2\overline{w}}+\frac{6\Gamma_{\max}(d-2,1)n^{2\gamma}\log n}{2^{k-1}n\overline{w}}\right)\sum_{l=1}^{K}\frac{(1+2\delta)^2}{2^{(l-1)(3-\beta)}}\\
&\leq\frac{\kappa^{2d-1}n^{2\gamma(3-\beta)(d-2)}n^{\gamma(7-\beta)}}{2^{k+1}C^2n^2}+\frac{6\kappa^{d-1}n^{\gamma((3-\beta)(d-2)+2)}\log n}{2^{k+1}Cn}=\nu_1,
\end{aligned}$$

where (a) holds due to the union bound; (b) follows from the union bound and event Ω^*_{d-1}.

Then, the event $\{y_i^k=1\}\setminus\{z_i^k=1\}$ denotes the event that i connects to some vertex in $\Gamma^{\overline{G}_1}_{d-1,k}(u)\setminus\Gamma^{\overline{G}_2}_{d-1,k}(v)$ and connects to some vertex in $\Gamma^{\overline{G}_2}_{d-1,k}(v)$ independently. Thus, $\mathbb{P}\left\{\{y_i^k=1\}\setminus\{z_i^k=1\}\mid\Omega^*_{d-1}\right\}$ can be bounded by

$$\begin{aligned}
&\mathbb{P}\left\{\{y_i^k=1\}\setminus\{z_i^k=1\}\mid\Omega^*_{d-1}\right\}\\
\leq&\mathbb{P}\left\{\exists j\in\Gamma^{\overline{G}_1}_{d-1,k}(u)\setminus\Gamma^{\overline{G}_2}_{d-1,k}(v):i\in\Gamma_1^{\overline{G}_1}(j)\mid\Omega^*_{d-1}\right\}\mathbb{P}\left\{\exists j\in\Gamma^{\overline{G}_2}_{d-1}(v):i\in\Gamma_1^{\overline{G}_2}(j)\mid\Omega^*_{d-1}\right\}\\
\leq&\mathbb{P}\left\{i\in\Gamma_d^{\overline{G}_1}(v)\mid\Omega^*_{d-1}\right\}\mathbb{P}\left\{i\in\Gamma_d^{\overline{G}_2}(v)\mid\Omega^*_{d-1}\right\}\\
\overset{(a)}{\leq}&\left(\frac{\kappa^d n^{\gamma((3-\beta)(d-1)+2)}}{2^{k+1}Cn}\right)^2\\
\leq&\frac{\kappa^{2d}n^{2\gamma(3-\beta)(d-2)}}{2^{2(k+1)}C^2n^2}n^{2\gamma(5-\beta)}=\nu_2,
\end{aligned}$$

where (a) follows from a similar proof of (41).

When we compare the first term of ν_1 and ν_2, we have

$$\frac{\kappa^{2d}n^{2\gamma(3-\beta)(d-2)}}{2^{2(k+1)}C^2n^2}n^{\gamma(7-\beta)}\leq\frac{\kappa^{2d}n^{2\gamma(3-\beta)(d-2)}}{2^{2(k+1)}C^2n^2}n^{2\gamma(5-\beta)},$$

where the last inequality follows from $\frac{n^{\gamma(7-\beta)}}{n^{2\gamma(5-\beta)}}=n^{\gamma(\beta-3)}\leq 1$.

Thus, we have

$$\begin{aligned}
\mathbb{P}\left\{y_i^k=1\mid\Omega^*_{d-1}\right\}&\leq\nu_1+\nu_2\leq 2\nu_2+\frac{4\kappa^d n^{\gamma((3-\beta)(d-2)+2)}\log n}{2^kCn}\\
&\leq\frac{\kappa^{2d}n^{2\gamma(3-\beta)(d-1)}n^{4\gamma}}{3\cdot 2^{2(k-1)}C^2n^2}+\frac{4\kappa^d n^{\gamma((3-\beta)(d-2)+2)}\log n}{2^kCn}\triangleq\mu_k.
\end{aligned}$$

Thus, conditional on Ω^*_{d-1}, we have

$$\left|\Gamma^{\overline{G}_1}_{d,k}(u)\cap\Gamma^{\overline{G}_2}_{d,k}(v)\right|\overset{s.t.}{\leq}\text{Binom}\left(\left|\overline{P}_k\right|,\mu_k\right).$$

Recall $n_k=Cn\alpha_{k-1}^{1-\beta}$ in view of (28) and $\left|\overline{P}_k\right|\leq 2n_k$ in view of (30). Therefore, for sufficiently large n,

$$2n_k\mu_k=\frac{2\kappa^{2d}n^{2\gamma(3-\beta)(d-1)}n^{\gamma(5-\beta)}}{3\cdot 2^{(k-1)(3-\beta)}Cn}+\frac{4\kappa^{d-1}n^{\gamma(3-\beta)(d-1)}\log n}{2^{(k-1)(2-\beta)}}\leq\frac{2}{3}\Psi(d,k).$$

We then apply Chernoff Bound with $\eta=\frac{1}{2}$ and get

$$\mathbb{P}\left\{\left|\Gamma^{\overline{G}_1}_{d,k}(u)\cap\Gamma^{\overline{G}_2}_{d,k}(v)\right|\geq\Psi_{\max}(d,k)\mid\Omega^*_{d-1}\right\}\leq n^{-4}.$$

Induction: Finally, we prove (46) by induction.
For $d = 1$, we have proved that (46) holds.
Suppose (46) holds for $d-1$, then taking the union bound yields that

$$\begin{aligned}\mathbb{P}\left\{\Omega_{d-1}^{*c}\right\} \leq & 2\cdot\mathbb{P}\left\{\Omega_{d-1}^{c}\right\} + \mathbb{P}\left\{\left|\Gamma_{d-1,k}^{\overline{G}_1}(u)\cap\Gamma_{d-1,k}^{\overline{G}_2}(v)\right| \geq \Psi_{\max}(d-1,k)\right\}\\ \leq & 2(4^{d-1}-1)n^{-4} + \frac{2\cdot 4^{d-1}}{3}n^{-4} = \left(\frac{2\cdot 4^d}{3}-1\right)\cdot n^{-4}.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}_1}(u)\cap\Gamma_{d,k}^{\overline{G}_2}(v)\right| \geq \Psi_{\max}(d,k)\right\}\\ \leq & \mathbb{P}\left\{\Omega_{d-1}^{*c}\right\} + \mathbb{P}\left\{\left|\Gamma_{d,k}^{\overline{G}_1}(u)\cap\Gamma_{d,k}^{\overline{G}_2}(v)\right| \geq \Psi_{\max}(d,k) \mid \Omega_{d-1}^{*}\right\}\\ \leq & \left(\frac{2\cdot 4^d}{3}-1\right)\cdot n^{-4} + n^{-4} \leq \frac{2\cdot 4^d}{3}\cdot n^{-4}.\end{aligned}$$

□

C.2.4 Proof of Lemma 4. The main idea of the proof is to bound the number of D-hop witnesses for both true pairs and fake pairs in the first slice, using the bounds to the number of the D-hop neighbors established in Lemma 2 and Lemma 3.

Recall that in Algorithm 1, we select the set $\widehat{\mathcal{S}}$ of low-degree seeds. Let $\widehat{S} = \{i : (i,i) \in \widehat{\mathcal{S}}\}$. To circumvent the dependency between $\widehat{S}$ and the graphs G_1 and G_2, we will introduce $\underline{S}$ and $\overline{S}$ such that they are independent from graphs and $\underline{S} \subset \widehat{S} \subset \overline{S}$ with high probability. To this end, we define an event $\mathcal{E}$ such that

$$\{i : w_i \leq c\} \subset \{i : |\Gamma_1^{G_1}(i)| \leq 5\log n, |\Gamma_1^{G_2}(i)| \leq 5\log n\} \subset \{i : w_i \leq \frac{15}{s}\log n\}.$$

For any i with $w_i \leq c$, $\mathbb{E}\left[|\Gamma_1^{G_1}(i)|\right] = cs$. Thus, applying Lemma 6 with $\lambda = 3\log n$ yields

$$\mathbb{P}\left\{|\Gamma_1^{G_1}(i)| \geq 5\log n\right\} \leq \mathbb{P}\left\{|\Gamma_1^{G_i}(i)| \geq 2cs + 4\log n\right\} \leq n^{-3}.$$

Taking a union bound over i gives $\mathbb{P}\left\{\{i : w_i \leq c\} \subset \{i : |\Gamma_1^{G_1}(i)| \leq 5\log n, |\Gamma_1^{G_2}(i)| \leq 5\log n\}\right\} \geq 1 - n^{-2+o(1)}$.

For any i with $w_i > \frac{15}{s}\log n$, $\mathbb{E}\left[|\Gamma_1^{G_1}(i)|\right] = 15\log n$. we apply Chernoff Bound in Theorem 3 with $\eta = 2/3$ and have

$$\mathbb{P}\left\{|\Gamma_1^{G_1}(i)| \leq 5\log n\right\} \leq \mathbb{P}\left\{|\Gamma_1^{G_i}(i)| \leq \left(1-\frac{2}{3}\right)15\log n\right\} \leq n^{-3}.$$

Thus, we have

$$\begin{aligned}&\mathbb{P}\left\{\{i : |\Gamma_1^{G_1}(i)| \leq 5\log n, |\Gamma_1^{G_2}(i)| \leq 5\log n\} \subset \{i : w_i \leq \frac{15}{s}\log n\}\right\}\\ =&\mathbb{P}\left\{\{i : w_i > \frac{15}{s}\log n\} \subset \{i : |\Gamma_1^{G_1}(i)| > 5\log n, |\Gamma_1^{G_2}(i)| > 5\log n\}\right\} = 1 - n^{-2+o(1)}.\end{aligned}$$

Thus, $\mathbb{P}\{\mathcal{E}\} \geq 1 - n^{-2+o(1)}$. On event $\mathcal{E}$, we have

$$\underline{S} \triangleq \{i : w_i \leq c\} \cap S \subset \widehat{S} \subset \{i : w_i \leq \frac{15}{s}\log n\} \cap S \triangleq \overline{S},$$

where $S = \{i : (i,i) \in \mathcal{S}\}$ denotes the set of vertices selected as the initial seed set $\mathcal{S}$. Note that crucially the initial seeds in $\mathcal{S}$ are selected among all true pairs with probability θ, independently from everything else. Thus $\underline{S}$ and $\overline{S}$ are independent from graphs. As a consequence, to bound from below (resp. above) the number of D-hop witnesses for the true (resp. fake) pair, it suffices to consider their common D-hop neighbors in $\underline{S}$ (resp. $\overline{S}$).

More specifically, let us first consider the true pairs. Fix any vertex $u \in P_1$. Let $\Lambda(u) = \Gamma_D^{G_1}(u) \cap \Gamma_D^{G_2}(u) \setminus \left(N_{D-1}^{\overline{G}_1}(u) \cap N_{D-1}^{\overline{G}_2}(u)\right)$. Define event

$$\mathcal{A}_u = \left\{\left|\Lambda(u) \cap \underline{S}\right| > \frac{3}{5}\Gamma_{\min}\theta\right\}, \text{ where } \Gamma_{\min} = \frac{1}{2}\left(\frac{C \cdot s^2}{12 \cdot \overline{w}}\right)^D n^{\gamma((3-\beta)(D-1)+1)}.$$

Note that due to assumption (10) and $n^{\gamma(3-\beta)} \gg \log n$ for sufficiently large n, $N_{\max} \le \frac{1}{10}\Gamma_{\min}$. Hence it follows from Lemma 2 that

$$\mathbb{P}\left\{|\Lambda(u) \cap \{i : w_i \le c| < \frac{4}{5}\Gamma_{\min}\right\} \le n^{-4+o(1)}.$$

Because the seeds $\mathcal{S}$ are selected among all true pairs with probability θ, independently from everything else, we have

$$\left|\Lambda(u) \cap \underline{S}\right| \sim \text{Binom}\left(|\Lambda(u) \cap \{i : w_i \le c\}|, \theta\right).$$

Then, we apply Chernoff Bound in Theorem 3 with $\eta = \frac{1}{4}$ and get

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{A}_u^c\right\} \le& \mathbb{P}\left\{|\Lambda(u) \cap \{i : w_i \le c\}| < \frac{4}{5}\Gamma_{\min}\right\} + \mathbb{P}\left\{\mathcal{A}_u^c \,\middle|\, |\Lambda(u) \cap \{i : w_i \le c\}| \ge \frac{4}{5}\Gamma_{\min}\right\} \\
\le& n^{-4+o(1)} + \mathbb{P}\left\{\text{Binom}\left(\Gamma_{\min}, \theta\right) \le \frac{3}{5}\Gamma_{\min}\theta\right\} \\
\le& n^{-4+o(1)} + \exp\left(-\frac{1}{40}\Gamma_{\min}\theta\right) \overset{(a)}{\le} n^{-4+o(1)},
\end{aligned}$$

where (a) holds due to assumption (11). Let $\mathcal{A} = \cap_{u \in P_1}\mathcal{A}_u$. It follows from the union bound that $\mathbb{P}\{\mathcal{A}\} \le n^{-3+o(1)}$.

We next consider the fake pairs. Fix any two distinct vertices $u, v \in \overline{P}_1$. Define an event

$$\mathcal{B}_{uv} = \left\{\left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \cap \overline{S}\right| \le \frac{1}{2}\Gamma_{\min}\theta\right\}.$$

Note that due to the assumption (10) and $n^{\gamma(3-\beta)} \gg \log^2 n$ for sufficiently large n,

$$\Psi_{\max} \le \frac{\kappa^{2D}}{\left(\frac{Cs^2}{12\cdot\overline{w}}\right)^D}\left(\frac{n^{\gamma((3-\beta)(D-1)+1)}}{Cn}\left(\frac{15}{s}\log n\right)^{3-\beta} + \frac{4+6\log n}{n^{\gamma(3-\beta)}}\right)\Gamma_{\min} \le \frac{1}{8}\Gamma_{\min}.$$

Hence, it follows from Lemma 3 that

$$\mathbb{P}\left\{\left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \cap \{i : w_i \le \frac{15}{s}\log n\}\right| > \frac{1}{4}\Gamma_{\min}\right\} \le n^{-4+o(1)}.$$

Since the seeds $\mathcal{S}$ are selected among all true pairs with probability θ independently, we have

$$\left|\Gamma_D^{\overline{G}_1}(u) \cap \Gamma_D^{\overline{G}_2}(v) \cap \overline{S}\right| \sim \text{Binom}\left(\left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \cap \{i : w_i \le \frac{15}{s}\log n\}\right|, \theta\right).$$

Then, we apply Chernoff Bound in Theorem 3 with $\eta = 1$ and get

$$\begin{aligned}\mathbb{P}\left\{\mathcal{B}_{uv}^c\right\} \le& \mathbb{P}\left\{\left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v) \cap \{i : w_i \le \frac{15}{s}\log n\}\right| > \frac{1}{4}\Gamma_{\min}\right\} \\ &+ \mathbb{P}\left\{\mathcal{E}_{uv}^c \,\middle|\, \left|N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(u) \cap \{i : w_i \le \frac{15}{s}\log n\}\right| \le \frac{1}{4}\Gamma_{\min}\right\} \\ \le& n^{-4+o(1)} + \mathbb{P}\left\{\mathrm{Binom}\left(\frac{1}{4}\Gamma_{\min}, \theta\right) \le \frac{1}{2}\Gamma_{\min}\theta\right\} \\ \le& n^{-4+o(1)} + \exp\left(-\frac{1}{12}\Gamma_{\min}\theta\right) \overset{(a)}{\le} n^{-4+o(1)},\end{aligned}$$

where (a) holds due to assumption (11). Let $\mathcal{B} = \cap_{u,v\in\overline{P}_1:u\neq v}\mathcal{B}_{uv}$. It follows from the union bound that $\mathbb{P}\{\mathcal{B}^c\} \le n^{-2+o(1)}$.

Finally, we define event C such that

$$\underline{G}_j \subset \widehat{G}_j \subset \overline{G}_j, \quad \forall j = 1, 2 \quad \text{and} \quad P_1 \subset \widehat{P}_1 \subset \overline{P}_1.$$

It follows from Lemma 1 that $\mathbb{P}\{C\} \ge 1 - n^{-4+o(1)}$. Taking the union bound, we have

$$\mathbb{P}\{\mathcal{A} \cap \mathcal{B} \cap C \cap \mathcal{E}\} \ge 1 - n^{-3+o(1)} - n^{-2+o(1)} - 2n^{-4+o(1)} \ge 1 - n^{-2+o(1)}.$$

It remains to verify that on the event $\mathcal{A} \cap \mathcal{B} \cap C \cap \mathcal{E}$, $\mathcal{R}_1$ contains all true pairs in Q_1 and no fake pairs in $\widehat{Q}_1$.

Recall that we uses seeds in $\widehat{\mathcal{S}}$ and count the D-hop witnesses in $\widehat{G}_1$ and $\widehat{G}_2$ for all candidate vertex-pairs in $\widehat{Q}_1$ in Step 4 of Algorithm 1. On event $\mathcal{A} \cap C \cap \mathcal{E}$, $\Lambda(u) \subset \Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(u)$ and the minimum number of D-hop witnesses among all true pairs (u, u) in Q_1 is lower bounded by $\frac{3}{5}\Gamma_{\min}\theta$. On event $\mathcal{B} \cap C \cap \mathcal{E}$, $\Gamma_D^{\widehat{G}_1}(u) \cap \Gamma_D^{\widehat{G}_2}(v) \subset N_D^{\overline{G}_1}(u) \cap N_D^{\overline{G}_2}(v)$ the maximum number of D-hop witnesses among all fake pairs (u, v) in $\widehat{Q}_1$ is upper bounded by $\frac{1}{2}\Gamma_{\min}\theta$. Thus, GMWM with threshold $\tau_1 = \frac{1}{2}\Gamma_{\min}\theta$ outputs $\mathcal{R}_1$, which contains all true pairs in Q_1 and no fake pairs in $\widehat{Q}_1$.

C.2.5 Proof of Lemma 5. Fix a vertex $u \in P_k$. For each vertex $i \in P_{k-1}$, let x_i be a binary random variable such that $x_i = 1$ if i connects to u both in G_1 and G_2, and $x_i = 0$ otherwise. Then, $\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k-1}\right| = \sum_{i\in P_{k-1}} x_i$ and x_i's are independent. Moreover, we have

$$\mathbb{P}\{x_i = 1\} \ge \frac{\alpha_k\alpha_{k-1}}{n\overline{w}}s^2.$$

Therefore, applying Chernoff Bound in Theorem 3 with $\eta = \frac{1}{3}$ yields that

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k-1}\right| \le \frac{C\alpha_{k-1}^{3-\beta}s^2}{3\overline{w}}\right\} \le \mathbb{P}\left\{\mathrm{Binom}\left(n_{k-1}, \frac{\alpha_k\alpha_{k-1}}{n\overline{w}}s^2\right) \le \frac{C\alpha_{k-1}^{3-\beta}s^2}{3\overline{w}}\right\} \le n^{-4},$$

where the last inequality holds because $n_{k-1}\frac{\alpha_k\alpha_{k-1}}{n\overline{w}}s^2 = \frac{C\alpha_{k-1}^{3-\beta}s^2}{2\overline{w}} \ge 32\log n$ in view of $(\alpha_{k^*})^{3-\beta} \ge \frac{85\overline{w}\log n}{Cs^2}$.

C.2.6 Proof of Lemma 6. Fix a pair of two distinct vertices $u, v \in \overline{P}_k$. For each vertex $i \in \overline{P}_{k-1}$, let x_i be a binary random variable such that $x_i = 1$ if i is connected to u in G_1 and v in G_2, and $x_i = 0$ otherwise. Since the event that i is connected to u is independent of the event that i is connected to

v, we have

$$\mathbb{P}\{x_i = 1\} \le \left((1+2\delta)^2 \frac{\alpha_{k-1}\alpha_{k-2}}{n\overline{w}} s\right)^2 = \frac{4(1+2\delta)^4 \alpha_{k-1}^4 s^2}{n^2 \overline{w}^2} \triangleq p_{\max}.$$

Moreover, x_i's are independent. Therefore, $\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{k-1}\right| \overset{s.t.}{\le} \text{Binom}\left(\left|\overline{P}_k\right|, p_{\max}\right)$. Recall $n_k = Cn\alpha_{k-1}^{1-\beta}$ in view of (28) and $\left|\overline{P}_k\right| \le 2n_k$ in view of (30). Thus, we apply Lemma 6 with $\lambda = 4\log n$, and get

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{k-1}\right| \ge \frac{16(1+2\delta)^4 C\alpha_{k-1}^{5-\beta} s^2}{\overline{w}^2 n} + \frac{16}{3}\log n\right\} \le n^{-4}.$$

C.2.7 Proof of Lemma 7. The proof is divided into two parts. The first part is to identify a set of "good" events whose intersection holds with high probability. The second part provides a deterministic argument, showing that on the intersection of these good events, the 1-hop algorithm successfully matches slice k for all $2 \le k \le k^*$.

First, we identify a good event under which the number of common 1-hop neighbors of a true pair is large. More precisely, for any vertex $u \in P_k$, define event

$$\mathcal{A}_k(u) = \left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k-1}\right| \ge \xi_k\right\}, \quad \text{where } \xi_k \triangleq \frac{C\alpha_{k-1}^{3-\beta} s^2}{3\overline{w}},$$

and $\mathcal{A} = \cap_{2\le k\le k^*} \cap_{u\in P_k} \mathcal{A}_k(u)$. By Lemma 5 and union bound, we have $\mathbb{P}\{\mathcal{A}^c\} \le n^{-3+o(1)}$.

Second, we determine a good event under which the number of common 1-hop neighbors of a fake pair is small. More formally, for any pair of distinct vertices $u, v \in \overline{P}_k$, define event

$$\mathcal{B}_k(u,v) = \left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{k-1}\right| \le \zeta_k\right\}, \quad \text{where } \zeta_k \triangleq \frac{16(1+2\delta)^4 C\alpha_{k-1}^{5-\beta} s^2}{\overline{w}^2 n} + \frac{16}{3}\log n,$$

and $\mathcal{B} = \cap_{2\le k\le k^*} \cap_{u,v\in\overline{P}_k : u\ne v} \mathcal{B}_k(u,v)$. By Lemma 6 and union bound, we have $\mathbb{P}\{\mathcal{B}^c\} \le n^{-3+o(1)}$.

Third, we define an event $\mathcal{C} = \cap_{2\le k\le k^*} \left\{Q_k \subset \widehat{Q}_k \subset \overline{Q}_k\right\}$. By Lemma 1 and union bound, we have $\mathbb{P}\{\mathcal{C}^c\} \le n^{-4+o(1)}$.

Finally, we let $\mathcal{F}$ denote the event that the first slice is successfully matched, i.e., $\mathcal{R}_1$ contains all true pairs in Q_1 and no fake pairs in $\widehat{Q}_1$. By Lemma 4, $\mathbb{P}\{\mathcal{F}^c\} \le n^{-1.5+o(1)}$.

Combining the above, it follows that

$$\mathbb{P}\{\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap\mathcal{F}\} \ge 1 - 2n^{-3+o(1)} - n^{-4+o(1)} - n^{-1.5+o(1)} \ge 1 - n^{-1.5+o(1)}.$$

It remains to verify on the event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap F$, $\mathcal{R}_k$ contains all true pairs in Q_k and no fake pairs in $\widehat{Q}_k$ for all $1 \le k \le k^*$. We prove this by induction. The base case with $k = 1$ follows from the definition of $\mathcal{F}$. Assume the induction hypothesis holds for the slice $k-1$, we aim to show it continues to hold for k.

Recall that when matching the slice $\widehat{Q}_k$, we use $\mathcal{R}_{k-1}$ as the set of seeds. Since the induction hypothesis is true for slice $k-1$, it follows that $\mathcal{R}_{k-1}$ contains all the true pairs in Q_{k-1}. Thus, the minimum number of 1-hop witnesses among all true pairs (u,u) in Q_k is lower bounded by ξ_k. Moreover, since $\mathcal{R}_{k-1}$ contains no fake pairs in $\widehat{Q}_{k-1}$ and on event $\mathcal{C}$, $\widehat{Q}_{k-1} \subset \overline{Q}_k$, it follows that $\mathcal{R}_{k-1}$ is contained by all the true pairs in $\overline{Q}_{k-1}$. Also, the set of fake pairs in $\widehat{Q}_k$ is contained by the set of fake pairs in $\overline{Q}_k$. Thus, the maximum number of 1-hop witnesses among all fake pairs (u,v) in $\widehat{Q}_k$ is upper bounded by ζ_k.

Note that

$$\xi_k \overset{(a)}{\geq} \tau_2(k) \text{ and } \frac{\zeta_k}{\tau_2(k)} \overset{(b)}{\leq} \frac{64(1+2\delta)^4 n^{2\gamma}}{\overline{w}n} + \frac{1}{3} \overset{(c)}{<} 1,$$

where (a) holds by definition of $\tau_2(k)$ in (5); (b) follows from $n^\gamma \geq \alpha_k \geq \alpha_{k^*} \geq \left(\frac{85\overline{w}\log n}{Cs^2}\right)^{\frac{1}{3-\beta}}$ for $2 \leq k \leq k^*$; (c) holds as n is sufficiently large in view of $n^{2\gamma} = o(n)$ and $\overline{w} = \Theta(1)$. Thus, $\mathcal{R}_k$ contains all true pairs in Q_k and no fake pairs in $\widehat{Q}_k$, completing the induction.

C.2.8 Proof of Lemma 8. Fix any two distinct vertices $u, v \in \overline{P}_{\geq k^*+1}$. Then $w_u, w_v \leq (1+2\delta)\alpha_{k^*}$. For each vertex $i \in \overline{P}_{\geq k^*}$, let x_i be a binary random variable such that $x_i = 1$ if i connects to u in G_1 and v in G_2, and $x_i = 0$ otherwise. Since the event that i connects to u is independent of the event that i connects to v, we have

$$\mathbb{P}\{x_i = 1\} \leq \left((1+2\delta)^2 \frac{\alpha_{k^*}\alpha_{k^*-1}}{n\overline{w}} s\right)^2 = (1+2\delta)^4 \frac{4\alpha_{k^*}^4 s^2}{n^2\overline{w}^2} \triangleq p_{\max}.$$

Moreover, x_i's are independent. Therefore,

$$\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{\geq k^*}\right| \overset{s.t.}{\leq} \text{Binom}\left(\left|\overline{P}_{\geq k^*}\right|, p_{\max}\right) \overset{s.t.}{\leq} \text{Binom}\left(n, p_{\max}\right).$$

Thus, we get

$$\begin{aligned}\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{\geq k^*}\right| \geq 3\right\} \leq& \mathbb{P}\left\{\text{Binom}\left(n, p_{\max}\right) \geq 3\right\} \\ \overset{(a)}{\leq}& n^3 p_{\max}^3 \\ \leq& \frac{(1+2\delta)^{12} C^3 \alpha_{k^*}^{12} s^6}{n^3 \overline{w}^6} \leq n^{-3+o(1)}.\end{aligned}$$

where (a) follows from the union bound.

C.2.9 Proof of Lemma 9. We first bound $|S_h|$ by conditioning on S_{h-1}. For any $u \in P_{k^*+h}$, let x_i be a binary random variable such that $x_i = 1$ if $i \in S_{h-1}$ connects to u, and $x_i = 0$ otherwise. Since S_{h-1} is only determined by the vertex weight and the edges connecting to previous S_l, $l < h-1$, the event that i and u is connected is independent across i conditional on S_{h-1}. It follows that

$$\mathbb{P}\{x_i = 1 \mid S_{h-1}\} \geq \frac{\alpha_{k^*+h+1}\alpha_{k^*+h}}{n\overline{w}} s^2.$$

Thus, we have $\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap S_{h-1}\right| \overset{s.t.}{\geq} \text{Binom}\left(|S_{h-1}|, \frac{\alpha_{k^*+h+1}\alpha_{k^*+h}}{n\overline{w}} s^2\right)$ conditional on S_{h-1}. Applying Chernoff Bound in Theorem 3 yields that

$$\begin{aligned}&\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap S_{h-1}\right| < 3 \,\middle|\, |S_{h-1}| \geq \frac{1}{2} n_{k^*+h-1}\right\} \\ \leq& \mathbb{P}\left\{\text{Binom}\left(\frac{1}{2} n_{k^*+h-1}, \frac{\alpha_{k^*+h+1}\alpha_{k^*+h}}{n\overline{w}} s^2\right) \leq (1-\eta)\mu\right\} \\ \leq& \exp\left(-\frac{\eta^2}{2}\mu\right) \triangleq p_h \leq \frac{1}{4},\end{aligned}$$

where $\mu = \frac{1}{2} n_{k^*+h-1} \frac{\alpha_{k^*+h+1}\alpha_{k^*+h}}{n\overline{w}} s^2 = \frac{C\alpha_{k^*+h}^{3-\beta} s^2}{4\overline{w}} \geq 16\log 2$ due to $\alpha_{k^*+h} \geq \left(\frac{64\overline{w}\ln 2}{Cs^2}\right)^{1/(3-\beta)}$ and $\eta = \frac{\mu-3}{\mu} \geq \frac{1}{2}$.

Then, the above result implies that: $\mathbb{E}\left[|S_h| \mid |S_{h-1}| \geq \frac{1}{2} n_{k^*+h-1}\right] \geq (1-p_h) n_{k^*+h}$. Note that the event $u \in S_h$ only depends on the vertex weight and the edge set $E_u \triangleq \{(u,i) : i \in S_{h-1}\}$. Because

E_u's are disjoint, the event $u \in S_h$ is independent across $u \in P_{k^*+h}$. Thus, we apply Chernoff Bound in Theorem 3 with $\eta = \frac{1-2p_h}{2(1-p_h)}$ and have

$$\mathbb{P}\left\{|S_h| < \frac{1}{2}n_{k^*+h} \,\middle|\, |S_{h-1}| \geq \frac{1}{2}n_{k^*+h-1}\right\} \leq \mathbb{P}\left\{\text{Binom}\left(n_{k^*+h}, 1-p_h\right) < \frac{1}{2}n_{k^*+h}\right\}$$
$$\leq \exp\left(-\frac{(1-2p_h)^2 n_{k^*+h}}{8(1-p_h)}\right) \leq n^{-3},$$

where the last inequality holds due to $n_{k^*+h} \geq n_{k^*} \geq Cn\left(\frac{85\overline{w}\log n}{Cs^2}\right)^{\frac{1-\beta}{3-\beta}} \geq 72\log n$ due to the choice of k^* in (4) and sufficiently large n.

Finally, we prove by induction that $\mathbb{P}\left\{|S_h| < \frac{1}{2}n_{k^*+h}\right\} \leq h \cdot n^{-3}$.

For $h = 0$, it is true by definition.

For $h \geq 1$, if $\mathbb{P}\left\{|S_{h-1}| \geq \frac{1}{2}n_{k^*+h-1}\right\} \geq 1-(h-1)\cdot n^{-3}$, then

$$\mathbb{P}\left\{|S_h| < \frac{1}{2}n_{k^*+h}\right\} \leq \mathbb{P}\left\{|S_h| < \frac{1}{2}n_{k^*+h} \mid |S_{h-1}| \geq \frac{1}{2}n_{k^*+h-1}\right\} + \mathbb{P}\left\{|S_{h-1}| < \frac{1}{2}n_{k^*+h-1}\right\}$$
$$\leq n^{-3} + (h-1)\cdot n^{-3} = h \cdot n^{-3}.$$

C.2.10 Proof of Lemma 10. First, for any two distinct vertices $u, v \in \overline{P}_{\geq k^*}$, define event

$$\mathcal{A}_{uv} = \left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{\geq k^*}\right| \leq 2\right\},$$

and $\mathcal{A} = \bigcap_{u,v\in\overline{P}_{\geq k^*}:u\neq v} \mathcal{A}_{uv}$. By Lemma 12 and union bound, we have $\mathbb{P}\{\mathcal{A}^c\} \leq n^{-1+o(1)}$.

Second, let $\mathcal{B}$ denote the event that all true pairs in P_{k^*} are matched successfully. By Lemma 7, $\mathbb{P}\{\mathcal{B}\} \geq 1 - n^{-1.5+o(1)}$.

Third, by Lemma 1 and union bound, we have $\mathbb{P}\left\{Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}\right\} \leq n^{-4+o(1)}$.

Finally, by Lemma 9, we have

$$\mathbb{P}\left\{|S_{h^*}| \geq \frac{1}{2}n_{k^*+h^*}\right\} \geq 1 - \cdot n^{-3+o(1)}.$$

Combining the above, it follows that

$$\mathbb{P}\left\{\mathcal{A} \cap \mathcal{B} \cap \{Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}\} \cap \{|S_{h^*}| \geq \frac{1}{2}n_{k^*+h^*}\}\right\} \geq 1 - n^{-1+o(1)}.$$

Now, suppose event $\mathcal{A} \cap \mathcal{B} \cap \{Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}\} \cap \{|S_{h^*}| \geq \frac{1}{2}n_{k^*+h^*}\}$ holds. We aim to show that $\mathcal{R}_{k^*+1}$ contains no fake pair in $\widehat{Q}_{\geq k^*}$ and all true pairs (u,u) with $u \in S_h$ for $h \geq 0$.

We first show $\mathcal{R}_{k^*+1}$ contains no fake pair in $\widehat{Q}_{\geq k^*}$. Suppose not. Let (u,v) denote the first fake pair in $\widehat{Q}_{\geq k^*}$ matched by the PGM algorithm. This implies that the PGM only matches true pairs before matching (u,v). Since the threshold r of the PGM is set to be 3, it follows that (u,v) has at least three 1-hop witnesses that are true pairs in $\widehat{Q}_{\geq k^*}$. Since $\widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}$, it follows that $\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{P}_{\geq k^*}\right| \geq 3$, which contradicts the fact that event $\mathcal{A}$ holds. Thus, $\mathcal{R}_{k^*+1}$ contains no fake pairs in $\widehat{Q}_{\geq k^*}$.

Next, we prove that $\mathcal{R}_{k^*+1}$ contains all true pairs in S_h for all $h \geq 0$ by induction. For ease of presentation, we assume $\mathcal{R}_{k^*+1}$ contains the match pairs in the previous slice, that is $\mathcal{R}_{k^*+1} \supset \mathcal{R}_{k^*}$. The base case with $h = 0$ follows from the definition of $\mathcal{B}$. Assume the induction hypothesis holds for $h-1$, we aim to show it continues to hold for h. Based on the definition of S_h, the true pairs in S_h have at least 3 common 1-hop neighbors in S_{h-1}. Because all true pairs in S_{h-1} have been

matched and $Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*}$, the true pairs in S_h would be matched by the PGM algorithm with threshold $r = 3$. Therefore, $\mathcal{R}_{k^*+1}$ contains all true pairs in S_h for all $h \geq 0$.

Finally,

$$|S_{h^*}| \geq \frac{1}{2} n_{k^*+h^*} = \frac{C}{2} n \alpha_{k^*+h^*}^{1-\beta} \geq \frac{C}{2} n (2\widetilde{w})^{1-\beta},$$

where $\widetilde{w} = \left(\frac{64\overline{w}\ln 2}{Cs^2}\right)^{1/(3-\beta)} = \Theta(1)$ and the last inequality holds due to the choice of h^*. Thus, $\mathcal{R}_{k^*+1}$ has $\Theta(n)$ true pairs.

C.2.11 Proof of Lemma 11. Fix a vertex $u \in P_0$. For each vertex $i \in P_{k^*}$, let x_i be a binary random variable such that $x_i = 1$ if i connects to u both in G_1 and G_2, and $x_i = 0$ otherwise. Then, $\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k^*}\right| = \sum_{i \in P_{k^*}} x_i$ and x_i's are independent. Moreover, we have

$$\mathbb{P}\{x_i = 1\} \geq \frac{\alpha_{k^*}\alpha_0}{n\overline{w}} s^2.$$

Recall $|P_{k^*}| \geq n_{k^*} = Cn\alpha_{k^*-1}^{1-\beta}$ in view of (28). Hence,

$$\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k^*}\right| \overset{s.t.}{\geq} \text{Binom}\left(n_{k^*}, \frac{\alpha_0\alpha_{k^*}}{n\overline{w}} s^2\right).$$

Thus, we apply Chernoff Bound in Theorem 3 with $\eta = \frac{1}{2}$ and get

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k^*}\right| \leq \frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{2\overline{w}}\right\} \leq \mathbb{P}\left\{\text{Binom}\left(n_{k^*}, \frac{\alpha_0\alpha_{k^*}}{n\overline{w}} s^2\right) \leq \frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{2\overline{w}}\right\} \leq n^{-4},$$

where the last inequality holds because $n_{k^*}\frac{\alpha_{k^*}\alpha_0}{n\overline{w}} s^2 = \frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{\overline{w}} \geq 64\log n$, due to the choice of k^* in (4).

C.2.12 Proof of Lemma 12. Fix two distinct vertices $u, v \in \overline{P}_0$. We bound from above the number of their common 1-hop neighbors in $\overline{R} = \bigcup_{k \geq 1} \overline{P}_k$.

For each $k \geq 1$ and each vertex $i \in \overline{P}_k$, let y_i^k be a binary random variable such that $y_i^k = 1$ if i is connected to u in G_1 and v in G_2, and $y_i^k = 0$ otherwise. Since the event that i is connected to u is independent of the event that i is connected to v, we have

$$\mathbb{P}\left\{y_i^k = 1\right\} \leq \left(\frac{(1+2\delta)\alpha_{k-1}w_{\max}}{n\overline{w}} s\right)^2 \leq (1+2\delta)^2 \frac{\alpha_{k-1}^2}{n\overline{w}} s^2 \triangleq p_{\max}^k, \quad \forall k \geq 1.$$

Moreover, y_i^k's are independent. Thus,

$$\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{R}\right| \overset{s.t.}{\leq} \sum_{k=1}^{K} \text{Binom}\left(\left|\overline{P}_k\right|, p_{\max}^k\right).$$

Recall $n_k = Cn\alpha_{k-1}^{1-\beta}$ in view of (28), $n_k \leq \left|\overline{P}_k\right| \leq 2n_k$, and $\kappa = \frac{(1+2\delta)^2 2^{5-\beta} C}{(2^{3-\beta}-1)\overline{w}}$. Thus,

$$\sum_{k=1}^{K} \left|\overline{P}_k\right| p_{\max}^k \leq \sum_{k=1}^{K} 2n_k \frac{(1+2\delta)^2\alpha_{k-1}^2}{n\overline{w}} s^2 = \frac{2Cn^{\gamma(3-\beta)}}{\overline{w}} \sum_{k=1}^{K} \frac{(1+2\delta)^2}{2^{(k-1)(3-\beta)}} \leq 3\kappa n^{\gamma(3-\beta)} s^2,$$

$$\sum_{k=1}^{K} \left|\overline{P}_k\right| p_{\max}^k \geq n_1 \frac{\alpha_0^2}{n\overline{w}} s^2 = \frac{Cn^{\gamma(3-\beta)}}{\overline{w}} s^2 \geq 64\log n.$$

Then, we apply Chernoff Bound in Theorem 3 with $\eta = \frac{1}{3}$, and get

$$\mathbb{P}\left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{R}\right| \geq 4\kappa n^{\gamma(3-\beta)}s^2\right\} \leq \mathbb{P}\left\{\sum_{k=1}^{K} \text{Binom}\left(\left|\overline{P}_k\right|, p_{\max}^k\right) \geq 4\kappa n^{\gamma(3-\beta)}s^2\right\} \leq n^{-4}.$$

C.2.13 Proof of Lemma 13. Recall the bound of the number of 1-hop witnesses is provided by Lemma 11 and Lemma 12.

First, for any vertex $u \in P_0$, define event

$$\mathcal{A}_u = \left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(u) \cap P_{k^*}\right| \geq \frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{2\overline{w}}\right\},$$

and $\mathcal{A} = \bigcap_{u\in P_0} \mathcal{A}_u$. By Lemma 11 and union bound, we have $\mathbb{P}\{\mathcal{A}\} \leq n^{-3+o(1)}$.

Second, for any two distinct vertices $u, v \in \overline{P}_0$, define event

$$\mathcal{B}_{uv} = \left\{\left|\Gamma_1^{G_1}(u) \cap \Gamma_1^{G_2}(v) \cap \overline{R}\right| \leq 4\kappa n^{\gamma(3-\beta)}s^2\right\},$$

and $\mathcal{B} = \bigcap_{u,v\in\overline{P}_0:u\neq v} \mathcal{B}_{uv}$. By Lemma 12 and union bound, we have $\mathbb{P}\{\mathcal{B}^c\} \leq n^{-2+o(1)}$.

Third, we define an event $C = \bigcap_{0\leq k\leq k^*}\left\{Q_k \subset \widehat{Q}_k \subset \overline{Q}_k\right\} \cap \left\{Q_{\geq k^*} \subset \widehat{Q}_{\geq k^*} \subset \overline{Q}_{\geq k^*}\right\}$. By Lemma 1 and union bound, we have $\mathbb{P}\{C^c\} \leq n^{-4+o(1)}$.

Finally, we let $\mathcal{E}$ denote the event that $\widehat{\mathcal{R}}$ contains all true pairs in Q_{k^*} and no fake pairs in $\widehat{Q}_k$ for any $k \geq 1$. By Lemma 4, Lemma 7 and Lemma 10, $\mathbb{P}\{\mathcal{E}^c\} \leq n^{-1.5+o(1)}$.

Combining the above, it follows that

$$\mathbb{P}\{\mathcal{A}\cap\mathcal{B}\cap C\cap\mathcal{E}\} \geq 1 - n^{-3+o(1)} - n^{-2+o(1)} - n^{-4+o(1)} - n^{-1.5+o(1)} \geq 1 - n^{-1.5+o(1)}.$$

Suppose $\mathcal{A}\cap\mathcal{B}\cap C\cap\mathcal{E}$ holds. Then $\widehat{\mathcal{R}}$ contains all true pairs in Q_{k^*}, and thus the minimum number of 1-hop witnesses among all true pairs (u,u) in $Q_0 \subset \widehat{Q}_0$ is lower bounded by $\frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{2n\overline{w}}$. Moreover, since $\widehat{\mathcal{R}}$ contains no fake pairs in $\widehat{Q}_{\geq 1}$ and $\widehat{Q}_{\geq 1} \subset \overline{Q}_{\geq 1}$ on event C, it follows that $\widehat{\mathcal{R}}$ is contained by all the true pairs in $\bigcup_{k\geq 1}\overline{Q}_k$, i.e., all the true pairs with weights no larger than $(1+2\delta)n^\gamma$. Thus, the maximum number of 1-hop witnesses among all fake pairs (u,v) in $\widehat{Q}_0 \subset \overline{Q}_0$ is upper bounded by $4\kappa n^{\gamma(3-\beta)}s^2$. Note that by the choice of k^* in (4), $\frac{C\alpha_{k^*}^{2-\beta}\alpha_0 s^2}{2\overline{w}} \geq \frac{Cn^\gamma s^2}{2\overline{w}}\left(\frac{85\overline{w}\log n}{Cs^2}\right)^{\frac{2-\beta}{3-\beta}} > 4\kappa n^{\gamma(3-\beta)}s^2$, where the last inequality hols for all sufficiently large n in view of $2 < \beta < 3$. Moreover, since $\overline{P_0} \subset P_0 \cup P_1$, for any fake pair $(u,v) \in \widehat{Q}_0$, the two corresponding true pairs $(u,u), (v,v) \in Q_0 \cup Q_1$. Therefore, the two true pairs either have more 1-hop witnesses than the fake pair (u,v) or have already been matched in $\widehat{Q}_1$. Hence, $\mathcal{R}_0$ contains all true pairs in Q_0 and no fake pairs in $\widehat{Q}_0$.

Received February 2021; revised April 2021; accepted April 2021