Efficient Algorithms for Exact Graph Matching on Correlated Stochastic Block Models with Constant Correlation

Joonhyuk Yang [*1] **Dongpil Shin** [*1] **Hye Won Chung** [1]

Abstract

We consider the problem of graph matching, or learning vertex correspondence, between two correlated stochastic block models (SBMs). The graph matching problem arises in various fields, including computer vision, natural language processing and bioinformatics, and in particular, matching graphs with inherent community structure has significance related to de-anonymization of correlated social networks. Compared to the correlated Erdős-Rényi (ER) model, where various efficient algorithms have been developed, among which a few algorithms have been proven to achieve the exact matching with constant edge correlation, no low-order polynomial algorithm has been known to achieve exact matching for the correlated SBMs with constant correlation. In this work, we propose an efficient algorithm for matching graphs with community structure, based on the comparison between partition trees rooted from each vertex, by extending the idea of Mao et al. (2021a) to graphs with communities. The partition tree divides the large neighborhoods of each vertex into disjoint subsets using their edge statistics to different communities. Our algorithm is the first low-order polynomial-time algorithm achieving exact matching between two correlated SBMs with high probability in dense graphs.

1. Introduction

Graph matching aims to align two (or more) graphs to reveal a bijection between the vertex sets such that the number of aligned edges is maximized. Given two graphs with n vertices, graph matching finds a solution for a quadratic assignment problem (QAP), $\max_{\Pi \in S_n} \langle A, \Pi B \Pi^\top \rangle$ over the set of all $n \times n$ permutation matrices S_n, where A and B denote the adjacency matrices of the two graphs. This problem has been studied in various fields, including social network analysis (Narayanan & Shmatikov, 2009), computer vision (Schellewald & Schnörr, 2005), pattern recognition (Conte et al., 2004), natural language processing (Haghighi et al., 2005), machine learning (Liu & Qiao, 2012), and bioinformatics (Kazemi et al., 2016; Chen & Yuan, 2006).

Although the QAP is known to be NP-hard in the worst case (Burkard et al., 1998), the graph matching can be solved in polynomial time for the average case of random graph models. Thus, many previous works have studied the graph matching for random graph models, especially for Erdős-Rényi (ER) graphs. In particular, the correlated ER graph model proposed by Pedarsani & Grossglauser (2011) has been widely studied. In this model, there exists a parent ER graph $G_0 \sim \mathcal{G}(n, p/(1-\alpha))$, and two subgraphs G and G' are obtained by independent subsampling of G_0, where each edge of G_0 is removed independently with probability α. The parameter α indicates the noise level, and $1-\alpha$ is the correlation between G and G'. Assuming a permutation $\pi : [n] \to [n]$ and denoting by G^π the graph obtained by permuting the vertices of G with π, exact graph matching aims to recover π from the two graphs $G^\pi, G' \sim \mathcal{G}(n, p)$.

For the correlated ER model, many works have focused on two fundamental questions: 1) deriving the information-theoretic limit on (n, p, α) for exact matching, and 2) developing polynomial-time algorithms for recovering π. Regarding the first question, it was shown in (Cullina & Kiyavash, 2016; Wu et al., 2022) that exact matching is achievable if $np(1-\alpha) \geq (1+\epsilon)\log n$ for any $\epsilon > 0$ where $p/(1-\alpha) = o(1)$. This limit implies that graph matching is information-theoretically possible for constant α if $np = \Theta(\log n)$ and even for α close to 1 if $np \gg \log n$. However, achieving this limit with polynomial-time algorithms is still open, although various efficient algorithms have been proposed (Ding et al., 2021; Fan et al., 2020; Mao et al., 2021b). Most of these algorithms require the noise level α to be arbitrarily close to 0 in order to guarantee exact matching. Some recent work has developed polynomial-time algorithms for constant correlation in the spare regime (Mao et al., 2021a) or in the sparse and dense regimes but with α less than some specific constant (Mao et al., 2022).

[*]Equal contribution [1]School of Electrical Engineering, KAIST, Daejeon, Korea. Correspondence to: Hye Won Chung <hwchung@kaist.ac.kr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

While this previous line of work has revealed several important aspects in matching correlated random graphs, it is still largely open how these results generalize to more practical random graph models beyond the Erdős-Rényi (ER) model. In particular, matching graphs with inherent community structure has relevance to real-world applications such as de-anonymization of social networks, but the investigation of efficient graph-matching algorithms for the graphs with community structure has been largely unexplored.

In this paper, we consider exact graph matching for random graphs with community structure, and develop a polynomial-time matching algorithm for constant edge-correlation. We focus on the correlated stochastic block models (SBMs), where the parent graph G_0 is assumed to be sampled from the SBM, which is known to be one of the most natural generative models for networks with community structure. To the best of our knowledge, our algorithm is the first low-order polynomial-time algorithm that guarantees exact matching of the correlated SBMs with constant correlation.

1.1. Correlated Stochastic Block Models

Consider an undirected graph of n vertices with a planted partition. Suppose the vertex set $[n]$ is partitioned into disjoint subsets of $k \geq 1$ communities, $C_1, C_2, \ldots, C_k$, where the number of vertices in community C_i is $|C_i| = n_i$ and $\sum_{i=1}^{k} n_i = n$. Without loss of generality, we assume that $n_1 \geq n_2 \geq \cdots \geq n_k := n_{\min}$. Given the partition $\{C_i\}_{i=1}^{k}$, the correlated stochastic block models (Onaran et al., 2016) are parameterized by $p, q \in [0,1]$, $p > q$, and $\alpha \in [0,1)$. We assume that there exists a parent graph G_0 with the given partition $\{C_i\}_{i=1}^{k}$, where the edges between each pair of vertices are drawn independently as follows: $u \in C_i$ and $v \in C_j$ are connected with probability $p/(1-\alpha)$ if $i = j \in [k]$ and with probability $q/(1-\alpha)$ otherwise. Then two subgraphs G and G' are obtained by independent subsampling of G_0 as follows: G is obtained by removing each edge of G_0 independently with probability α, and G' is obtained independently in the same way as G. Note that $\mathbb{P}\{(i,j) \in \mathcal{E}(G) | (i,j) \in \mathcal{E}(G')\} = 1 - \alpha$ where $\mathcal{E}(G)$ is the edge set of the graph G. Given a permutation $\pi : [n] \to [n]$, G^{π} is the graph obtained by permuting the vertices of G by π. Our goal is to design a polynomial-time algorithm that can exactly recover the permutation π at the constant noise level α. We will prove the performance guarantees of the proposed algorithm in terms of the parameters $(p, q, \alpha, k, n_{\min})$, both for the cases with and without the knowledge of the exact community structure $\{C_i\}_{i=1}^{k}$.

1.2. Prior Work and Main Question

Graph matching on correlated stochastic block models has been studied from two different aspects: first, when the community structure is revealed in both G^{π} and G', what are the information-theoretically feasible regimes where exact matching is possible so that the vertex identities can be de-anonymized; second, when the goal is to recover the community structure of the original graph G_0 from the subsampled graph G', what is the regime where having G^{π} as side information can be beneficial? The first aspect was studied in Onaran et al. (2016); Cullina et al. (2016) for the two-community case ($k = 2$), and the sufficient condition for the optimal maximum likelihood estimator to recover π with high probability is shown to be $(1-\alpha)(p+q)n/2 > 2(\log n)$ when $n_1 = n_2 = n/2$. The second aspect was studied in (Racz & Sridhar, 2021) for the case of two equal-sized communities, and it was shown that even if the communities of G^{π} and G' are not revealed, the maximum likelihood estimator can recover π when $(1-\alpha)(p+q)n/2 > (\log n)$, which matches the impossibility results for the exact matching proved in (Cullina et al., 2016). Note that $(1-\alpha)(p+q)n/2 > (\log n)$ is the necessary condition to guarantee that there are no isolated vertices in $G \cap G'$ with high probability. The result of Racz & Sridhar (2021) implies that there exist regimes of (p, q, α) where having the correlated graph G^{π} can bring an information advantage in recovering the community of G' (or G_0) through the graph matching between G^{π} and G'.

These previous works establish the information-theoretic thresholds for exact graph matching of the correlated SBMs with two communities, and in particular, it is shown that when $p = \Theta(q) = \Omega(\log n/n)$, graph matching is information-theoretically possible with constant correlation $(1-\alpha)$, even if the community structure is not revealed in both graphs. However, no polynomial-time algorithm has been proven to guarantee exact graph matching for the correlated SBMs with constant correlation.

When the partition of vertices based on their community labels is revealed in both G^{π} and G', it is possible to apply the graph matching algorithm for the correlated ER model to each of the communities, since the subgraphs induced by C_i in G^{π} and G' are the correlated ER graphs with a parent graph $G_0(C_i) \sim \mathcal{G}(n_i, p/(1-\alpha))$. In Table 1, we summarize the results of the previous algorithms, originally developed for the correlated ER model, when applied to each recovered community $(C_1, \cdots, C_k)$ of the correlated SBMs. Here we simply assume that all the communities have the same size of $m := n/k$.

The 'Black swan' algorithm (Barak et al., 2019) allows the largest noise level of $\alpha = 1 - (\log m)^{-o(1)}$, but it requires quasi-polynomial time complexity $m^{\Theta(\log m)}$. This algorithm defines a rich family of rare subgraphs called 'black swans', each consisting of $\Theta(\log m)$ vertices, and finds the rare subgraphs that appear in both graphs with $m^{\Theta(\log m)}$ time complexity. The 'Chandelier' (Mao et al., 2022) algorithm also uses the idea of 'subgraph counting', but with a

Table 1: Comparison with literature for the exact matching of the correlated SBMs with k-balanced communities of size m.

Algorithm	Density	Noise level	Time complexity
Black swan (Barak et al., 2019)	$mp \geq m^{o(1)}$	$\alpha \leq 1 - (\log m)^{-o(1)}$	$k \times m^{O(\log m)}$
Degree profile (Ding et al., 2021)	$mp \geq (\log m)^C$	$\alpha \leq (\log m)^{-C}$	$k \times O(m^3 p^2 + m^{2.5})$
GRAMPA (Fan et al., 2019)	$mp \geq (\log m)^C$	$\alpha \leq (\log m)^{-C}$	$k \times O(m^3)$
Binary tree (Mao et al., 2021a)	$(1+\epsilon)\log m \leq mp \leq m^{1/(C \log\log m)}$	$\alpha \leq \min\{\text{const.}, \epsilon/4\}$	$k \times m^{2+o(1)}$
Chandelier (Mao et al., 2022)	$(1+\epsilon)\log m \leq mp(1-\alpha)$	$\alpha \leq 1 - \sqrt{0.338}$	$\geq k \times m^{25}$
Our result ($2^{k'}$-ary partition tree)	$(\log m)^C \leq mp \leq m^{1/20}$ $mq = \Omega\left((\log\log m)^2\right)$	$\alpha \leq$ const.	$k \times O(m^4 p)$

family of unbalanced rooted trees of size $\Theta(\log m)$, called 'chandeliers', which can be counted approximately in polynomial time. This algorithm succeeds in graph matching for both sparse and dense regimes with constant correlation, but the time complexity scales in high-order polynomials, $\Omega(m^{25})$.

Among computationally more efficient methods with time complexity $O(m^3)$, the only algorithm that has been proven to guarantee exact matching with constant correlation is the 'Binary (partition) tree' algorithm by Mao et al. (2021a). The main idea is to define a signature vector of dimension $\Theta(\log m)$ associated with each vertex $i \in [m]$ by using the degree information of a large neighborhood of each i. By measuring the similarity between the signature vectors, the algorithm recovers the exact match for m vertices. The signature vector is defined by constructing a depth-ℓ binary partition tree rooted at each vertex $i \in [m]$, where at each depth $r \in [\ell]$ the neighborhood around the vertex i with radius-r is partitioned into 2^r disjoint subsets. To define a signature vector of dimension $2^\ell = \Theta(\log m)$, the depth of the binary partition tree needs to be $\ell = \Theta(\log\log m)$. Thus, the graph should be sparse as $mp \leq m^{1/(C \log\log m)}$ to avoid generating a loop in the partition tree, and this limits the application of this algorithm in the denser regime.

The main open question is how to construct an efficient matching algorithm for the correlated SBMs in the dense regime. We focus on the fact that the edges between communities have not been exploited in the previous algorithms. By using the correlation of both intra- and inter-community edges, as well as the known community structure, we develop a matching algorithm using the similarity between the signature vectors of the correct pairs, as in Mao et al. (2021a), but by constructing $2^{k'}$-ary (rather than binary) partition trees, which effectively reduces the required depth of the tree and thus successfully operates in the dense regime.

1.3. Main Results

Our main contribution is the development of a low-order polynomial-time algorithm for the exact matching of the correlated SBMs with constant correlation in the dense regime.

Theorem 1.1 (Exact matching for the correlated SBMs with known community structure)**.** *There exist absolute constants $\alpha_1, M, M' > 0$ with the properties below. Consider the two graphs G^π and G', which are generated from the correlated SBMs defined in Sec. 1.1, with the underlying permutation π and correlation $1-\alpha$. Suppose that the community labels $(C_1, C_2, \ldots, C_k)$ are given in both graphs, and assume that C_k is the smallest community of size $n_{\min}$. Assume that $\alpha \in (0, \alpha_1)$, $n_{\min} = \Omega(n^{10/19})$ and*

$$(\log n_{\min})^{1.1} \leq n_{\min} p \leq n_{\min}^{1/20}, \tag{1.1}$$

$$n_{\min} q \geq M'(\log\log n_{\min})^2, \tag{1.2}$$

$$k \geq \left(\frac{M(\log\log n_{\min}) \cdot (\log n_{\min} p)}{\log n_{\min}} \vee 3 \right). \tag{1.3}$$

Then, there exists a polynomial-time algorithm that recovers π exactly with high probability as $n \to \infty$. The complexity of the algorithm scales as $O(km^4 p)$ for the balanced communities of size $m := n/k$.

Even if the community structure is not revealed in both graphs, we can apply our matching algorithm by imposing an additional assumption on $n_{\min}$, which is sufficient for a polynomial-time algorithm to recover the community structure in both graphs. We combine Theorem 1.1 with the result of Yan et al. (2018), which proposed a semidefinite programming (SDP) for community detection with provable guarantees.

Corollary 1.2 (Exact matching for the correlated SBMs with unknown community structure)**.** *There exist absolute constants $\alpha_1, M, M', M_1 > 0$ with the properties below. Consider the two graphs G^π and G', which are generated from the correlated SBMs defined in Sec. 1.1, with the underlying permutation π and correlation $1-\alpha$. Assume that the sizes of all communities are different and the smallest community size is $n_{\min}$. Assume that $\alpha \in (0, \alpha_1/2)$ and (1.1), (1.2) and (1.3) hold. Also assume that*

$$n_{\min} \geq M_1 \left(\frac{\sqrt{np}}{p-q} \vee \frac{p \log n}{(p-q)^2} \right). \tag{1.4}$$

Then, there exists a polynomial-time algorithm that recovers π exactly with high probability as $n \to \infty$.

Remark 1.3. The assumption in Corollary 1.2 that community sizes are unique can be satisfied with high probability even if the community label of each vertex is sampled from a distribution, as long as each label is sampled with probability $\Theta(1/k)$. We prove this in Appendix D.1.

1.4. Notation

Asymptotic dependencies are denoted by the standard notations $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot), \Theta(\cdot)$ with $n \to \infty$. For $n \in \mathbb{N}^+$, we denote the set $[n] := \{1, 2, \dots, n\}$. Let $\wedge$ denote the minimum operator and $\vee$ denote the maximum operator. For a real number x, $\mathbf{Sign}(x) = 1$ if $x \geq 0$ and $\mathbf{Sign}(x) = -1$ if $x < 0$. For a graph G, we define $\mathcal{N}_G(i)$ as the set of neighbors of vertex i in G, and $\mathcal{N}_G(S)$ as the union of $\mathcal{N}_G(i)$ over all i in set $S \subset [n]$. For $d \in \mathbb{N}^+$, $\mathcal{B}_G(i, d)$ ($\mathcal{S}_G(i, d)$) denotes the ball (sphere) of radius d centered at vertex i in graph G. For a subset $R \subset [n]$ and a permutation $\pi : [n] \to [n]$, $G(R)$ denotes the subgraph of G induced by R, and $\pi|_R$ denotes the permutation over R defined by π. Let $\deg_G(i)$ denote the degree of i in a graph G. For a graph G with communities $(C_1, C_2, \dots, C_k)$, $\deg_G^a(i)$ denotes the size of neighbors of i in the community C_a.

2. Algorithm and Results

2.1. Overview of Algorithm

Consider the correlated SBMs (G^π, G') with correlation $1-\alpha$. Suppose that the community structure $(C_1, C_2, \dots, C_k)$ is revealed in both graphs, and $|C_1| \geq |C_2| \geq \cdots \geq |C_k| = n_{\min}$. Our algorithm consists of three stages. In the first stage, we focus on the minimum-size community C_k and generate a signature vector for each vertex in $G^\pi(C_k)$ and $G'(C_k)$. In this stage, we use both the edges within the community C_k and the edges across C_k and other (randomly sampled) $k' < k-1$ communities, say $C_1, \dots, C_{k'}$. Each vertex $i \in G^\pi(C_k)$ and $j \in G'(C_k)$ is associated with a signature vector $f(i) \in \mathbb{R}^{2^{k'\ell}}$ and $f'(j) \in \mathbb{R}^{2^{k'\ell}}$, respectively, where the $(2^{k'\cdot\ell})$-dimensional signature is obtained from a depth-ℓ $(2^{k'})$-ary partition tree rooted from i (j), which partitions the vertices in the sphere $\mathcal{S}_{G^\pi(G_k)}(i, r)$ ($\mathcal{S}_{G'(G_k)}(j, r)$) at each depth $r \in [\ell]$ into $2^{k'r}$ disjoint subsets. The detailed steps of generating the partition trees and the signature vectors are explained in Sec. 2.2. We will choose (k', ℓ) such that $k' \cdot \ell = \Theta(\log\log n_{\min})$ so that the signature contains enough information to distinguish and match $n_{\min}$ vertices in $G^\pi(G_k)$ and $G'(G_k)$. The maximum depth ℓ of the partition tree needs to satisfy $(n_{\min}p)^\ell = O(n_{\min})$ to guarantee that the partition tree does not contain a loop with high probability, and this gives a condition on k' such that $k' \geq \frac{C(\log\log n_{\min})\cdot(\log n_{\min}p)}{\log n_{\min}}$ for our algorithm to operate successfully.

In the second stage, we compare the signature vectors $f(i)$ and $f'(j)$ for each (i, j) pair in $G^\pi(C_k)$ and $G'(C_k)$, and find a pair whose "normalized distance" is less than some threshold. This gives an almost exact matching for the vertices in C_k, and a subsequent refinement step gives the exact matching for $n_{\min}$ vertices.

In the last stage, we use the matched vertex pairs in $G^\pi(C_k)$ and $G'(C_k)$ as initial seeds, and apply a seeded graph matching algorithm multiple times to match the rest of the vertices. In the following, we explain each of the three stages in more detail.

Algorithm 1 Vertex Signature Using Community Structure

Input: a graph Γ with community structure $(C_1, \dots, C_k)$ where $|C_k| = n_k$, parameters $k', \ell \in \mathbb{N}$, and $i \in C_k$

Output: $f(i) \in \mathbb{R}^{2^{k'\ell}}$ and $\mathrm{v}(i) \in \mathbb{R}^{2^{k'\ell}}$

$T_\varnothing^0 \leftarrow \{i\}$
for $r = 0, \dots, \ell-1$ **do**
  for $\boldsymbol{s}^r \in \{-1,1\}^{k'r}$ and $\boldsymbol{s}_{r+1} \in \{-1,1\}^{k'}$ **do**
    $T^{r+1}_{(\boldsymbol{s}^r, \boldsymbol{s}_{r+1})}(i) \leftarrow$
    $\{j \in \mathcal{N}_{\Gamma(C_k)}(T^r_{\boldsymbol{s}^r}(i)) \cap \mathcal{S}_{\Gamma(C_k)}(i, r+1) :$
    $\mathbf{Sign}(\deg_\Gamma^a(j) - n_a q) = \boldsymbol{s}_{r+1}(a)$ for $a \in [k']\}$
  end for
end for
Define $f(i), \mathrm{v}(i) \in \mathbb{R}^{2^{k'\ell}}$ by
$f_{\boldsymbol{s}^\ell}(i) := \sum_{j \in T^\ell_{\boldsymbol{s}^\ell}(i)} \left(\deg_\Gamma^k(j) - 1 - n_k p\right)$,
$\mathrm{v}_{\boldsymbol{s}^\ell}(i) := n_k p(1-p)\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right|$ for $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$.
return $f(i)$ and $\mathrm{v}(i)$

2.2. Partition Trees using Community Structure

In this section, we explain the process of constructing a partition tree for each vertex $i \in C_k$ in a graph $\Gamma \sim \mathrm{SBM}(n, p, q)$ using a known community structure $(C_1, \dots, C_k)$ for $[n]$. This partition tree is used to define a signature vector $f(i)$ for each $i \in C_k$. We randomly select $k' < k-1$ communities and denote them by $C_1, \dots, C_{k'}$ for notational simplicity. For a given vertex $i \in C_k$, the partition tree starts from a root $T_\varnothing^0 := \{i\}$, and at each level $r = 1, \dots, \ell$ the tree partitions the vertices in $\mathcal{S}_{\Gamma(C_k)}(i, r)$ into $2^{k'r}$ disjoint subsets, denoted by nodes $T^r_{(\boldsymbol{s}_1, \dots, \boldsymbol{s}_r)}$ for $(\boldsymbol{s}_1, \dots, \boldsymbol{s}_r) \in \{-1,1\}^{k'r}$. For notational simplicity, let $\boldsymbol{s}^r := (\boldsymbol{s}_1, \dots, \boldsymbol{s}_r)$. For every $r < \ell$, the node $T^r_{\boldsymbol{s}^r}$ has $2^{k'}$ children $T^{r+1}_{(\boldsymbol{s}^r, \boldsymbol{s}_{r+1})}$, $\boldsymbol{s}_{r+1} \in \{-1,1\}^{k'}$, where $T^{r+1}_{(\boldsymbol{s}^r, \boldsymbol{s}_{r+1})}$ contains all vertices $j \in \mathcal{N}_{\Gamma(C_k)}(T^r_{\boldsymbol{s}^r}) \cap \mathcal{S}_{\Gamma(C_k)}(i, r+1)$ satisfying $\mathbf{Sign}(\deg_\Gamma^a(j) - n_a q) = \boldsymbol{s}_{r+1}(a)$ for each $a \in [k']$, where $\boldsymbol{s}_{r+1}(a)$ is the a-th entry of $\boldsymbol{s}_{r+1} \in \{-1,1\}^{k'}$. In other words, we check whether the degree of vertex j to each community C_a, $a \in [k']$, is greater than or equal to the average degree $n_a q$, encode this information by $\boldsymbol{s}_{r+1} \in \{-1,1\}^{k'}$, and assign the vertex j to the corresponding node set $T^{r+1}_{(\boldsymbol{s}^r, \boldsymbol{s}_{r+1})}$ in the partition tree. After

constructing the partition tree of depth-ℓ rooted at i, the signature vector $f(i) \in \mathbb{R}^{2^{k'\ell}}$ is generated based on the degrees of the vertices in the leaves $T^{\ell}_{\boldsymbol{s}^\ell}$, $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$.We also define a variance vector $\mathrm{v}(i) \in \mathbb{R}^{2^{k'\ell}}$ that stores the variances of each entry of the signature vector $f(i)$.

Algorithm 1 describes the process of inductively constructing the partition tree and generating the signature vector.

2.3. Exact Matching for the Smallest Community

The second stage aims to match all the correct vertex pairs in $G^\pi(C_k)$ and $G'(C_k)$ by computing and comparing $(f(i), f'(j))$ for all possible pairs $i \in G^\pi(C_k)$ and $j \in G'(C_k)$. For this stage, we follow the two-step procedures, designed by Mao et al. (2021a), where the first-step finds a rough estimate of the permutation over $[n_k]$ vertices by comparing the similarity between the signature vectors, and the second step refines the output of the first step to recover the exact permutation. In the first step, every pair of $i \in G^\pi(C_k)$ and $j \in G'(C_k)$ is compared in terms of the normalized distance $\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(j)\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(j)}$ where J is a random subset uniformly drawn from $\{-1,1\}^{k'\ell}$ with polylogarithmic cardinality in n_k. If there exists (i,j) whose distance is less than the threshold $|J|(1 - \frac{1}{\sqrt{\log n_k}})$ then the pair is matched. To resolve the case where more than one vertices $j \in G'(C_k)$ are matched to the same vertex $i \in G^\pi(C_k)$ or no vertex is matched to $i \in G^\pi(C_k)$, a clean-up step is applied to generate a permutation $\tilde{\pi}_k : [n_k] \to [n_k]$ over C_k. Algorithm 2 describes the process of obtaining an initial estimate $\tilde{\pi}_k$ of the permutation. Given the estimate $\tilde{\pi}_k$, one can apply the refinement step to obtain the permutation $\hat{\pi}_k : [n_k] \to [n_k]$ equal to the true one $\pi|_{C_k}$. Due to space limitations, we present the refinement algorithm (Algorithm 4 (Mao et al., 2021a)) in Appendix §J.

Algorithm 2 Almost Exact Matching (Mao et al., 2021a)

Input: two graphs Γ and Γ' with community structure $(C_1, \ldots, C_k)$ where $|C_k| = n_k$
Output: a permutation $\tilde{\pi}_k : [n_k] \to [n_k]$

$\ell \leftarrow \left\lceil \frac{\log n_k}{40 \log n_k p} \right\rceil \wedge \lceil 42 \log\log n_k \rceil$
$w \leftarrow \lfloor (\log n_k)^5 \rfloor$, $k' \leftarrow \lceil 1680 (\log\log n_k) \frac{\log n_k p}{\log n_k} \rceil$
for $i \in C_k$ **do**
  $(f(i), \mathrm{v}(i)) \leftarrow$ Algorithm 1(Γ, k', ℓ, i)
  $(f'(i), \mathrm{v}'(i)) \leftarrow$ Algorithm 1(Γ', k', ℓ, i)
end for
$J \leftarrow$ a random subset uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$,
for $i \in \Gamma(C_k)$ and $j \in \Gamma'(C_k)$ **do**
  if $\sum_{\boldsymbol{s}^\ell \in J} \frac{(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(j))^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(j)} < 2w \left(1 - \frac{1}{\sqrt{\log n_k}}\right)$ **then**
    $B_{i,j} \leftarrow 1$
  else
    $B_{i,j} \leftarrow 0$
  end if
end for
$H \leftarrow$ the bipartite graph with adjacency matrix B
let $V = V' = [n_k]$
while H has at least one edge **do**
  choose a random edge $i \sim j$ in H for $(i,j) \in V \times V'$
  define $\tilde{\pi}(j) := i$ and remove the edge $i \sim j$ from H
  $V \leftarrow V \backslash \{i\}$, $V' \leftarrow V' \backslash \{j\}$
end while
if $V \neq \varnothing$ **then**
  define $\tilde{\pi}|_V$ as an arbitrary bijection from V to V'
end if
return $\tilde{\pi}_k$

Theorem 2.1 (Almost Exact Matching on C_k)**.** *For any constant $D > 0$ there exist constants n_0 and c depending on D, and absolute constants α_1, M, M' with the properties below. Consider the two graphs G^π and G', generated from the correlated SBMs defined in Sec. 1.1, with the underlying permutation $\pi : [n] \to [n]$. Suppose that community labels are given in both graphs, and C_k is the smallest community of size $n_{\min}$. Assume that $\alpha \in (0, \alpha_1)$, $n_{\min} \geq n_0$, (1.1), (1.2) and (1.3) hold. Then, Algorithm 2 applied to the inputs (G^π, G') with the known community structure in both graphs yields the permutation $\tilde{\pi}_k$ over C_k such that*

$$|i \in C_k : \tilde{\pi}_k(i) \neq \pi(i)| \leq 4 n_{\min}^{1-c}. \tag{2.1}$$

with probability at least $1 - n_{\min}^{-D}$.

By combining Theorem 2.1 with the result of Mao et al. (2021a), where it was shown that the exact matching of the correlated ER model can be obtained from the almost exact matching under some mild conditions on the graphs, which can already be satisfied by the conditions in Theorem 2.1, we can obtain the result for the exact matching.

Corollary 2.2 (Exact matching on C_k)**.** *Assume the conditions of Theorem 2.1. Let $\tilde{\pi}_k : [n_k] \to [n_k]$ be the output of Algorithm 2 when the inputs are the correlated SBMs G^π and G'. Then, when the refinement matching algorithm (Algorithm 4 (Mao et al., 2021a)) is applied to $\tilde{\pi}_k$, the output $\hat{\pi}_k : [n_k] \to [n_k]$ recovers the original permutation over C_k, i.e., $\hat{\pi}_k = \pi|_{C_k}$, with probability at least $1 - 2n_{\min}^{-D}$.*

Remark 2.3 (Conditions on k and q)**.** In Theorem 2.1, in addition to the density condition (1.1), we need two additional conditions on the number k of communities (1.3) and on the edge density q between communities (1.2). The condition on k is required to have the signature vector of dimension $2^{k'\ell} = \Theta(\log n_{\min})$ at the maximum tree depth ℓ, the diameter of C_k. If $n_{\min} p = O(n_{\min}^{1/R \log\log n_{\min}})$, we only need a constant number of communities, but in the denser regime $n_{\min} p = n_{\min}^c$, $c > 0$, we need $k = \Theta(\log\log n_{\min})$, since

Algorithm 3 Seeded Matching

Input: two graphs Γ and Γ' with community $(C_1, \dots, C_k)$, a permutation $\pi_s : [n_s] \to [n_s]$ over C_s
Output: a permutation $\hat{\pi}_t : [n_t] \to [n_t]$ over C_t

```
for i ∈ C_t and i' ∈ C_t do
    let the weight w(i, i') = 0
    for each j ∈ C_s do
        if ∃z ∈ N_{Γ(C_t)}(i) and ∃z' ∈ N_{Γ'(C_t)}(i') such that
        (z, π_s(j)) ∈ E(Γ) and (z', j) ∈ E(Γ') then
            w(i, i') = w(i, i') + 1
        end if
    end for
    if w(i, i') ≥ (n_t n_s pq)/8 then
        π̂_t(i') = i
    end if
end for
return π̂_t
```

the diameter is a constant. The condition on q is needed to guarantee that there are enough edges between C_k and other communities so that the $2^{k'}$-ary partition tree is well constructed by using the edges between communities.

2.4. Seeded Graph Matching

The last stage aims to recover the complete permutation $\pi : [n] \to [n]$ using the matched pairs in C_k as initial *seeds*. The main idea is to apply a seeded graph matching algorithm (similar to that in (Korula & Lattanzi, 2013)), where vertex pairs $i \in G^\pi(C_t)$ and $i' \in G'(C_t)$ are compared by counting the number of common seed pairs in the 2-hop neighborhoods of i and i'. Algorithm 3 describes the exact seeded matching algorithm, and Theorem 2.4 describes the condition for recovering a permutation π_t over C_t by using a known permutation π_s over C_s.

Theorem 2.4 (Seeded Matching for a Community)**.** *There exists an absolute constant $K > 0$ with the properties below. Consider the two graphs G^π and G', generated from the correlated SBMs defined in Sec. 1.1, with the underlying permutation $\pi : [n] \to [n]$. Suppose that the community labels $(C_1, C_2, \dots, C_k)$ are given in both graphs. Additionally, the exact permutation $\pi_s : [n_s] \to [n_s]$ for the vertices in community C_s is given. Assume that the parameters (p, q, n_t, n_s, α) of the model satisfy $\alpha < \frac{1}{10}$, $p \le \frac{1}{256}$, and*

$$n_t p \ge K \log n_t,\ n_t n_s pq \ge K \log n_t,\ n_t pq \le 1/256. \tag{2.2}$$

Then Algorithm 3 applied to G^π and G' yields $\hat{\pi}_t = \pi|_{C_t}$ with probability at least $1 - 4/n_t$.

Using the recovered permutation $\pi_k : [n_k] \to [n_k]$ over C_k, we first apply Algorithm 3 to recover the permutation over C_r for some $r \notin \{1, \dots, k', k\}$, which is a community not used in the previous stage of signature vector generation. We can check that all the conditions in (2.2) except the last one are satisfied by the conditions on $|C_k| := n_{\min}$ and (p, q) in (1.1) and (1.2), since $n_t \ge n_s = n_{\min}$. To satisfy $n_t pq \le 1/256$ for any $t \in \{1, \dots, k-1\}$, we need an additional condition on $n_{\min}$ such that $n_{\min} \ge M_1 n^{10/19}$. With these conditions, the exact matching for the vertices in C_r can be recovered by Algorithm 3. Then, given $\pi_r : [n_r] \to [n_r]$, we use the matched pairs over C_r as seeds and recover the matching over the rest of the communities $(C_1, \dots, C_{r-1}, C_{r+1}, \cdots, C_{k-1})$. The reason we do not use C_k but C_r as seeds in the permutation recovery is to avoid the dependency issue caused by reusing the edges over C_k and $(C_1, \dots, C_{k-1})$ that were used in generating the signature vectors for the vertices of C_k.

By combining Corollary 2.2 and Theorem 2.4, we obtain our complete result to exactly recover $\pi : [n] \to [n]$.

Corollary 2.5 (Exact Matching)**.** *For any constant $D > 0$ there exist a constant n_0 depending on D and absolute constants $\alpha_1, M, M', M_1 > 0$ with the properties below. Consider the two graphs G^π and G', generated from the correlated SBMs defined in Sec. 1.1, with the underlying permutation $\pi : [n] \to [n]$ and correlation $1 - \alpha$. Suppose that community labels $(C_1, C_2, \dots, C_k)$ are given in both graphs, and assume that C_k is the smallest community with size $n_{\min}$. Assume that $\alpha \in (0, \alpha_1)$, (1.1), (1.2) and (1.3) hold. Also assume that $n_{\min} \ge M_1 n^{10/19}$. Run Algorithm 2 with input graphs (G^π, G') to get an initial estimate of the permutation $\tilde{\pi}_k : [n_k] \to [n_k]$ over C_k. Using $\tilde{\pi}_k$ as input, run the refinement matching algorithm (Algorithm 4) to obtain $\hat{\pi}_k : [n_k] \to [n_k]$ over C_k. Using $\hat{\pi}_k$ as input, run Algorithm 3 to find a matching $\hat{\pi}_r : [n_r] \to [n_r]$ over C_r that was not used in Algorithm 2. Then, by using $\hat{\pi}_r$ as input, run Algorithm 3 repeatedly to recover $\hat{\pi}_i$, for all $i \in [k] \backslash \{r, k\}$. By combining the recovered permutations over each community, we get $\hat{\pi} := (\hat{\pi}_1, \dots, \hat{\pi}_k) : [n] \to [n]$ such that $\hat{\pi} = \pi$ with probability at least $1 - 2n_{\min}^{-D} - n^{-1/19}$.*

Note that Corollary 2.5 implies Theorem 1.1.

Remark 2.6 (Time complexity)**.** We summarize the total time complexity. For simplicity, we assume that the size of each community is equal to $m := n/k$. First, the time complexity to run Algorithm 1 is $O((mp)^\ell \cdot (k'mq))$, since the depth-$\ell$ partition tree contains $(mp)^\ell$ vertices, and to assign each vertex to one of the partitioning nodes at each level, $(k'mq)$ neighboring vertices need to be checked on average. Since we will choose $(mp)^\ell = O(m)$, this time-complexity is bounded by $O(k'm^2q)$. In Algorithm 2, we apply Algorithm 1 to m vertices in C_k and compare m^2 pairs of signature vectors of dimension $(\log m)^5$. Thus, the complexity is $O(k'm^3q + m^2(\log m)^5)$. In addition, the refinement matching by Algorithm 4 requires time complexity of $O(m^3)$. The seeded graph matching by Algorithm 3 requires $O(m^2(mp)(m))$-time complexity, and we run this

algorithm a total of $(k-1)$ times. Thus, the complexity for seeded matching is $O(km^4p)$. The total time complexity is dominated by that of the seeded graph matching, $O(km^4p)$.

Remark 2.7 (Density regime)**.** We discuss the most general density regime, where there exist low-order polynomial-time algorithms for exact matching of the correlated SBMs with constant correlation. Assume approximately balanced community sizes, i.e., $n_{\min} = \Theta(n/k)$. Our main result (Theorem 1.1) shows that exact matching is achievable in polynomial time for the correlated SBMs whose average degrees within the smallest community satisfy $(\log n_{\min})^{1.1} \le n_{\min}p \le n_{\min}^{1/20}$ with two additional conditions on the number k of communities (1.3) and the edge density q between communities (1.2). The result of Mao et al. (2021a) (summarized in Table 1), on the other hand, shows the existence of an efficient algorithm in the sparse regime $(1+\epsilon)\log n_{\min} \le n_{\min}p \le n_{\min}^{1/(C\log\log n_{\min})}$ for some constant $C>0$. For both the algorithms, we need an additional assumption on $n_{\min}$ in (1.4), which is equivalent to $k \le C'(\sqrt{\frac{n}{p}}(p-q) \wedge \frac{n(p-q)^2}{p\log n})$ for some constant $C'>0$, in order to recover the community labels in both graphs before performing graph matching. In summary, for the correlated SBMs with the density $(1+\epsilon)\log n \le np \le n^{1/20}$, there exists a low-order polynomial-time algorithm that can achieve exact matching of vertices if $\alpha \in (0, \text{const.})$ and $\frac{C\log\log n \cdot \log np}{\log n} \le k \le C'(\frac{\sqrt{n}(p-q)}{\sqrt{p}} \wedge \frac{n(p-q)^2}{p\log n})$. For the dense regime, we also need $nq/k = \Omega\left((\log\log n)^2\right)$.

3. Outline of Proof

In this section, we outline the proof of Theorem 2.1. Recall that Algorithm 1 constructs a $2^{k'}$-ary partition tree of depth ℓ, rooted from each vertex $i \in [C_k]$, of size $|C_k| = n_{\min}$, where the tree rooted from i is constructed within C_k by using the intra-community edges, while the partitioning of the vertices at each depth of the tree is done using the inter-community edges between C_k and $(C_1, \dots, C_{k'})$, the randomly selected k' communities. Our proof begins by showing that for $\ell = \lceil \frac{\log n_{\min}}{40\log n_{\min}p} \rceil \wedge \lceil 42\log\log n_{\min} \rceil$ and $k'\ell = \Theta(\log\log n_{\min})$, the length-$\ell$ neighborhoods of the majority of vertices in C_k form trees, and the vertices in the length-ℓ sphere $\mathcal{S}_{G(C_k)}(i,\ell)$ of i are uniformly partitioned into the leaf nodes $T^{\ell}_{\boldsymbol{s}^\ell}(i)$, $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$, each with a rough size of $(n_{\min}p/2^{k'})^\ell$.

We then compare the overlap between $T^{\ell}_{\boldsymbol{s}^\ell}(i,G)$ and $T^{\ell}_{\boldsymbol{s}^\ell}(i,G')$, the leaf nodes of the partition trees of $i \in C_k$ in graphs G and G', resp., and show that there is significant overlap, $\left|T^{\ell}_{\boldsymbol{s}^\ell}(i,G) \cap T^{\ell}_{\boldsymbol{s}^\ell}(i,G')\right| \ge (n_{\min}p/2^{k'})^\ell(1-\epsilon)^{k'\ell}$, where ϵ is an arbitrary small constant. This overlap leads to a correlation between the signature vectors $f(i)$ and $f'(i)$. When comparing the ℓ_2-normalized distance between $f(i)$ and $f'(i)$, we use the sparsification procedure suggested by Mao et al. (2021a), to mitigate the dependency in the entries of the signature vector of i caused by the overlap between the sets $T^{\ell}_{\boldsymbol{s}^\ell}(i,G)$ and $T^{\ell}_{\boldsymbol{t}^\ell}(i,G)$ for $\boldsymbol{s}^\ell \ne \boldsymbol{t}^\ell$. After sparsification, the normalized distance between the sparsified signatures of correct pairs is shown to be smaller than a given threshold with high probability.

For incorrect pairs of vertices $i \ne i'$, the sizes of $T^{\ell}_{\boldsymbol{s}^\ell}(i,G) \cap \mathcal{B}_{G_0(C_k)}(i',\ell)$ and $T^{\ell}_{\boldsymbol{s}^\ell}(i',G') \cap \mathcal{B}_{G_0(C_k)}(i,\ell)$ summed over the sparsified set are well controlled. As a result, the normalized distance between the sparsified signatures of distinct vertices is larger than the threshold with high probability. Detailed proofs for theorems are available in Appendix.

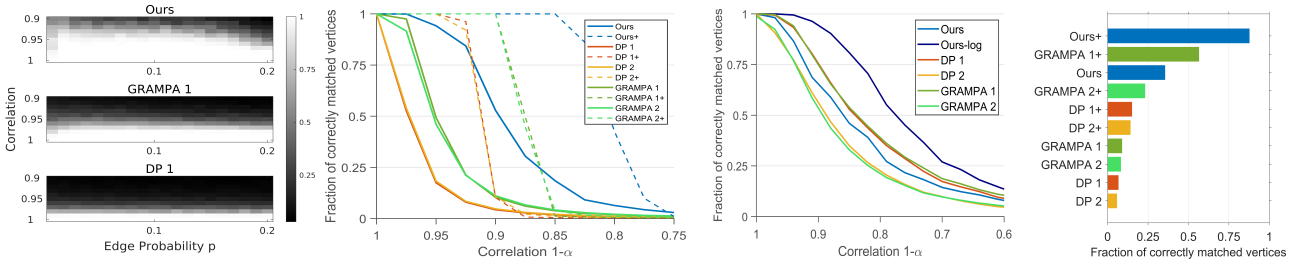
4. Experiment

In this section, we evaluate our algorithm on both synthetic and real networks with inherent community structure.

Datasets We use three types of datasets described in Table 2. For the synthetic dataset, we sample a parent graph G_0 from SBM with $p = 0.025$ and $q = p/3$, where the vertices are partitioned into six communities of equal size. Then, two graphs G and G' are subsampled independently from G_0 by removing each edge of G_0 with probability $\alpha \in [0,1)$. For real datasets, we obtain correlated graphs in two different ways: 1) For the 'BlogCatalog' dataset (Li et al., 2015), the parent graph is given by the network of the blogger community, where each user is treated as a vertex and the edges are connected between the users if they follow each other. We subsample the parent graph twice independently to generate the correlated graphs. The community label is given based on the predefined categories of the blogs. 2) For the 'Movie' dataset, there is no parent graph, but two correlated networks are constructed from two different movie datasets from 'Rotten Tomatoes' and 'IMDb' that share common vertices (movie items). Edges are connected between vertices if there is at least one common actor in the top cast list, where the 'Rotten Tomatoes' dataset contains up to six actors per movie and the 'IMDb' dataset contains up to five actors per movie. The community label is assigned based on six groups of release years. We also perform experiments on two more real datasets, Flickr data (Li et al., 2015) and ACM-DBLP data (Zhang & Tong, 2019), and present the results in Appendix A.5.

Table 2: Details of three networks.

Type	Model/Source	n	# of edges	k
Synthetic	SBM	4,998	139k	6
Sampled	BlogCatalog	5,196	171k	6
Correlated	Rotten-Tomatoes	4,780	110k	6
	IMDb	4,780	90k	6



(a) SBM varying p and $1-\alpha$ (b) SBM varying $1-\alpha$ (c) BlogCatalog varying $1-\alpha$ (d) Movie

Figure 1: Comparison of our partition tree algorithm with other baselines, GRAMPA and Degree Profile (DP) for matching networks of (a)-(b) correlated SBMs, (c) BlogCatalog network and (d) two correlated movie networks ('RT' and 'IMDb').

Baselines We compare the performance of our algorithm to two baselines: 'GRAMPA' (Fan et al., 2020) and 'Degree Profile (DP)' (Ding et al., 2021). As reviewed in Table 1, these algorithms are computationally efficient graph-matching algorithms with provable performance guarantees. GRAMPA is an algorithm based on a spectral method where a similarity matrix is constructed using all eigenvector pairs of the two graphs. Then, a linear assignment problem applied to the similarity matrix yields a match (permutation over the vertices). Degree Profile computes and compares the degree profile of the neighbors of each vertex, which is used to generate a similarity matrix. Since these baselines are developed in a way that does not use the community structure, for a fair comparison, we consider two variants of these algorithms (Methods 1 and 2): Method 1 applies the original algorithm (either GRAMPA or Degree Profile) to the whole graph of $[n]$ vertices to generate a $n \times n$ similarity matrix, but then selects a subpart of the matrix corresponding to the vertices in C_k to define the similarity matrix over $[n_k]$. After solving the assignment problem to generate a permutation $\tilde{\pi}_k : [n_k] \to [n_k]$, we apply the rest of the steps same as ours, including the refinement matching for C_k by Algorithm 4 (Mao et al., 2021a) and seeded graph matching (Algorithm 3) using the matched pairs in C_k as initial seeds. Method 2 applies the original algorithm to each of the recovered communities to obtain the similarity matrices, and then solves the assignment problem and refinement matching for each of the communities. To compare the performance of our algorithm with these baselines, we also define a $(n_k \times n_k)$-dimensional similarity matrix S with its (i,j)-th entry $S_{ij} := \sum_{\boldsymbol{s}^\ell} \frac{(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(j))^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(j)}$. Instead of thresholding the entries of S as in our Algorithm 2, we apply the linear assignment problem to S, as in GRAMPA (Fan et al., 2020), to compare the accuracy of the recovered permutation over C_k. We also compare the final accuracy of these three algorithms.

Our code is publicly available at `https://github.com/cabaksa/cSBM_Matching`.

Results We present the experimental results for the SBM networks in Fig. 1a–1b, the BlogCatalog networks in Fig. 1c and the Movie networks in Fig. 1d, respectively. Each baseline (GRAMPA and DP) has two versions (1 and 2), as explained before. The lines without $+$ indicate the fraction of correctly matched vertices within the initial comparison set C_k after solving the linear assignment problem on the similarity matrix of each algorithm, and those with $+$ indicate the final matching accuracy over the $[n]$ vertices.

In Fig. 1a, we plot the empirical performances of our algorithm, GRAMPA1, and Degree Profile1, averaged over 10 runs, as we change the intra-community edge density p ($q = p/3$) and the correlation parameter $1-\alpha$. A lighter color indicates a lower fractional error. We can observe that our algorithm is more robust against the decrease of the correlation $1-\alpha$ for all $p \in [0.01, 0.2]$ range. In Fig. 1b we can observe that our algorithm outperforms other baselines in recovering the permutation $\pi_k : [n_k] \to [n_k]$ over C_k (solid lines) and maintains high accuracy longer as the correlation decreases. The final matching accuracy (dotted lines) is also better for our algorithm.

The degree distribution of BlogCatalog graphs is known to follow a power law (Li et al., 2015). Since Algorithm 2 computes the normalized distance between the signature vectors assuming the binomial degree distribution, we adjust the definition of the signature vectors by using the log of the degree instead of the degree itself. This variation is denoted 'Ours-log' in Fig. 1c. We can observe that 'Our-log' outperforms other baselines, although the accuracy of our original algorithm without the degree fixing is lower than that of GRAMPA1 and DP1.

For 'Movie' networks, unlike other data, there is no parent graph, and the two correlated graphs have different edge densities. The common edges between the two graphs are ∼80k, which means that the correlation $1-\alpha$ is 0.73 and 0.89, respectively, from the perspective of each graph. Figure 1d shows the fraction of correctly matched vertices on this dataset, where our algorithm achieves the best performance.

5. Discussion and Open Problems

We presented a polynomial-time algorithm with low-order complexity for achieving exact matching on correlated SBMs with constant correlation. Our result is the first of its kind in the dense regime. In the sparse regime of the graphs where $(1+\epsilon)\log n_{\min} \le n_{\min}p \le n_{\min}^{1/(C\log\log n_{\min})}$, by directly applying the binary partition tree algorithm from Mao et al. (2021a) to each recovered community, one can achieve the exact graph matching with constant correlation in polynomial-time complexity. By generalizing the idea from Mao et al. (2021a), we developed a $2^{k'}$-ary partition tree algorithm for exact graph matching for the correlated SBMs, using the known community structure and the edges between communities. Our algorithm achieves the exact matching in $(\log n_{\min})^{1.1} \le n_{\min}p \le n_{\min}^{1/20}$ with constant correlation, with two additional conditions on the minimum number of communities k and the edge density across communities. Our work leaves several important open questions, as discussed below.

Exact graph matching without exact community recovery In the correlated SBMs, where G^π and G' are sampled from a parent graph G_0, one can try to match the vertices of the correlated graphs G^π and G' using a known community structure, or one can try to recover the community labels by first matching the two correlated graphs. The second problem was considered in (Racz & Sridhar, 2021) for the case of two equal-sized communities, and it was shown that there exist regimes of (p, q, α) where having the correlated graph G^π as side information can provide an information advantage in recovering the community of G' (or G_0) by graph matching between G^π and G'. Our work addressed the first problem and showed that there exists a low-order polynomial-time algorithm that can achieve the exact graph matching with constant correlation by using the known community structure in both graphs. However, it requires an additional assumption on $n_{\min}$, the minimum size of the community, as in (1.4), to recovery the community structure exactly in polynomial time. Then the natural question is whether exact community recovery is necessary to achieve exact graph matching with constant correlation in polynomial time. The answer may be no, and this is an interesting future work.

Exact graph matching via exact community recovery Even if we assume that the community labels in both graphs are given as side information, there is a gap between the information-theoretic limits for exact graph matching and the regime where the current polynomial algorithms can exactly recover the matching between the vertices of the correlated SBMs. First, our algorithm achieves the exact graph matching in the dense regime, but up to $n_{\min}p \le n_{\min}^{1/20}$. The upper bound on $n_{\min}$ is needed in the proof of Lemma F.6 to guarantee that the majority of vertices in C_k form trees up to depth ℓ. Thus, we may need another algorithmic approach beyond the partition tree to achieve the exact matching in the denser regime, e.g., $n_{\min}p = n_{\min}^{1-o(1)}$. For the sparse regime, to apply the algorithm from Mao et al. (2021a) to each community, we need the condition $(1+\epsilon)\log n_{\min} \le n_{\min}p \le n_{\min}^{1/(C\log\log n_{\min})}$. Consider the correlated SBMs with two balanced (size $n/2$ each) communities. The information-theoretic limit from Racz & Sridhar (2021) shows that exact matching is possible if $\frac{a+b}{2}(1-\alpha) > 1$ on the correlated SBMs with two balanced communities, where $p = \frac{a\log n}{(1-\alpha)n}$, and $q = \frac{b\log n}{(1-\alpha)n}$ for positive constant a, b. Thus, if $a = 1.5$, $b = 1$, and α is a sufficiently small constant, the exact matching is information-theoretically possible. However, since the density of each community in this regime is $n_{\min}p \approx 0.75\log n_{\min} < \log n_{\min}$, we cannot apply the matching algorithm from Mao et al. (2021a) to each community. Thus, there is a gap between the information-theoretic limit and the computational limit when we apply the matching algorithm to each recovered community. This implies that we need another algorithmic approach that uses both inter-/intra-community edges even in the sparse regime to bridge the gap between the information-theoretic limit and the computational limit.

Using seeded graph matching Our algorithm uses Algorithm 2 to achieve the almost exact matching on community C_k, and then uses the refinement matching (Algorithm 4) to achieve exact matching on community C_k. Finally, it uses vertex pairs from the community C_k as seeds to complete exact matching for the rest of the vertices in the correlated SBMs. In the correlated ER model, the seeded graph matching algorithm has been extensively studied. There are seeded graph matching algorithms based on percolation (Yartseva & Grossglauser, 2013; Chiasserini et al., 2014), and algorithms using large neighbor statistics (Mossel & Xu, 2020). Furthermore, Kazemi et al. (2015); Lubars & Srikant (2018); Yu et al. (2021) have proposed algorithms for seeded graph matching, even when the initial seeds are noisy. It is also worth investigating how we can improve the conditions for our exact graph matching algorithm by adjusting the first step and allowing some noisy matching from the initial seed set C_k.

6. Acknowledgements

This research was supported by the National Research Foundation of Korea under grant 2021R1C1C11008539, and by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2018-0-01402) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

Barak, B., Chou, C.-N., Lei, Z., Schramm, T., and Sheng, Y. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Chen, J. and Yuan, B. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

Chiasserini, C., Garetto, M., and Leonardi, E. De-anonymizing scale-free social networks by percolation graph matching. *CoRR*, abs/1411.7296, 2014.

Conte, D., Foggia, P., Sansone, C., and Vento, M. Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.*, 18(3):265–298, 2004.

Cullina, D. and Kiyavash, N. Improved achievability and converse bounds for erdos-rényi graph matching. *ACM SIGMETRICS performance evaluation review*, 44(1):63–72, 2016.

Cullina, D., Singhal, K., Kiyavash, N., and Mittal, P. On the simultaneous preservation of privacy and community structure in anonymized networks. *arXiv preprint arXiv:1603.08028*, 2016.

Ding, J., Ma, Z., Wu, Y., and Xu, J. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115, 2021.

Fan, Z., Mao, C., Wu, Y., and Xu, J. Spectral graph matching and regularized quadratic relaxations I: the gaussian model. *CoRR*, abs/1907.08880, 2019.

Fan, Z., Mao, C., Wu, Y., and Xu, J. Spectral graph matching and regularized quadratic relaxations: Algorithm and theory. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2985–2995. PMLR, 13–18 Jul 2020.

Fiori, M., Sprechmann, P., Vogelstein, J., Musé, P., and Sapiro, G. Robust multimodal graph matching: Sparse coding meets graph matching. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Kazemi, E., Hassani, S. H., and Grossglauser, M. Growing a graph matching from a handful of seeds. *Proceedings of the VLDB Endowment*, 8(10):1010–1021, 2015.

Kazemi, E., Hassani, S. H., Grossglauser, M., and Modarres, H. P. PROPER: global protein interaction network alignment through percolation matching. *BMC Bioinform.*, 17: 527:1–527:16, 2016.

Korula, N. and Lattanzi, S. An efficient reconciliation algorithm for social networks. *arXiv preprint arXiv:1307.1690*, 2013.

Li, J., Hu, X., Tang, J., and Liu, H. Unsupervised streaming feature selection in social media. CIKM '15, pp. 1041–1050. Association for Computing Machinery, 2015. ISBN 9781450337946.

Liu, Z.-Y. and Qiao, H. A convex-concave relaxation procedure based subgraph matching algorithm. In *Asian Conference on Machine Learning*, pp. 237–252. PMLR, 2012.

Lubars, J. and Srikant, R. Correcting the output of approximate graph matching algorithms. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1745–1753. IEEE, 2018.

Mao, C., Rudelson, M., and Tikhomirov, K. Exact matching of random graphs with constant correlation. *arXiv preprint arXiv:2110.05000*, 2021a.

Mao, C., Rudelson, M., and Tikhomirov, K. Random graph matching with improved noise robustness. In *Conference on Learning Theory*, pp. 3296–3329. PMLR, 2021b.

Mao, C., Wu, Y., Xu, J., and Yu, S. H. Random graph matching at otter's threshold via counting chandeliers. *arXiv preprint arXiv:2209.12313*, 2022.

Mossel, E. and Xu, J. Seeded graph matching via large neighborhood statistics. *Random Structures & Algorithms*, 57(3):570–611, 2020.

Narayanan, A. and Shmatikov, V. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pp. 173–187. IEEE, 2009.

Okamoto, M. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10(1):29–35, 1959.

Onaran, E., Garg, S., and Erkip, E. Optimal de-anonymization in random graphs with community structure. In *2016 50th Asilomar conference on signals, systems and computers*, pp. 709–713. IEEE, 2016.

Pedarsani, P. and Grossglauser, M. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1235–1243, 2011.

Racz, M. and Sridhar, A. Correlated stochastic block models: Exact graph matching with applications to recovering communities. *Advances in Neural Information Processing Systems*, 34:22259–22273, 2021.

Schellewald, C. and Schnörr, C. Probabilistic subgraph matching based on convex relaxation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 171–186. Springer, 2005.

Wu, Y., Xu, J., and Sophie, H. Y. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory*, 2022.

Yan, B., Sarkar, P., and Cheng, X. Provable estimation of the number of blocks in block models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1185–1194. PMLR, 2018.

Yartseva, L. and Grossglauser, M. On the performance of percolation graph matching. In *Proceedings of the first ACM conference on Online social networks*, pp. 119–130, 2013.

Yu, L., Xu, J., and Lin, X. Graph matching with partially-correct seeds. *J. Mach. Learn. Res.*, 22:280:1–280:54, 2021.

Zhang, S. and Tong, H. Attributed network alignment: Problem definitions and fast solutions. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1680–1692, 2019. doi: 10.1109/TKDE.2018.2866440.

The appendix of this paper is organized as follows. Additional experimental details and results are presented in Sec. A, and the proofs of the theoretical results follow. The proof of Corollary 1.2, where we give the sufficient conditions for exact matching of the correlated SBMs with unknown community structure, is presented in Sec. B. Theorem 2.4, which presents the performance of seeded graph matching (Algorithm 3), is proved in Sec. C. The proof of Corollary 2.5 is given in Sec. D. The proof of our main theorem (Theorem 2.1) is presented in Sec. E. Various lemmas to prove Theorem 2.1 are presented in Sections F, G and H. Some technical tools to prove the main results are summarized in Sec. I, and the previous result of Mao et al. (2021a) regarding the refinement algorithm for exact matching is summarized in Sec. J.

A. Experiment

A.1. Experimental Details

- Movie data: To generate two correlated networks sharing a common set of vertices, we use the 'Rotten Tomatoes'[1] dataset and the 'IMDb'[2] dataset. We generate two networks that share a common set of vertices corresponding to 4780 movies. Edges are connected between vertices if there is at least one common actor in the top cast list, where the 'Rotten Tomatoes' dataset contains up to six actors per movie and the 'IMDb' dataset contains up to five actors per movie. The community label is assigned based on six groups of release years.

- Community labels: In all the experiments reported in Section 4, to focus on the performance comparisons between different matching algorithms, we assume that the ground-truth community labels are given in the networks.

- Parameters: In order to apply our algorithm (Algorithm 1), the model parameters (p, q) are needed to obtain the values of $n_a q, n_k p$ and $n_k p(1-p)$, which are used to compute the signatures. Since these parameters are unknown in real datasets, we estimate these values from the data. Specifically, the medians of the vertex degree within and across communities are used instead of $n_k p$ and $n_a q$, respectively, and the empirical variance of vertex degree within C_k replaces $n_k p(1-p)$.

- Signature vector: The two parameters (k', ℓ), which determine the dimension of the signature vectors, can be considered as hyperparameters of our algorithm. In our experiments, we choose $k' = 4$ and $\ell = 2$. In Appendix §A.3, we provide additional experiments with varying (k', ℓ).

- Similarity matrix and assignment problem: The two main baselines (GRAMPA and Degree Profile) use two-step procedures, generating the similarity matrix over the vertices and solving the assignment problem to output a permutation from the similarity matrix. To compare the performance of our algorithm with these baselines, we also define a $(n_k \times n_k)$-dimensional similarity matrix S by defining its (i,j)-th entry as $S_{ij} := \sum_{\boldsymbol{s}^\ell} \frac{(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(j))^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(j)}$. If both the partitioning nodes $T^\ell_{\boldsymbol{s}^\ell}(i)$ and $T'^\ell_{\boldsymbol{s}^\ell}(j)$ are empty for some $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$, we set $\frac{(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(j))^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(j)} := 0$. Instead of thresholding the entries of S as in our Algorithm 2 to generate the permutation over C_k, we apply the linear assignment problem to S suggested in (Fan et al., 2020), to compare the accuracy of the permutation obtained from our similarity matrix with that of the two baselines.

- Seeded matching : In Algorithm 3, we count the number of common seeds on 2-hop neighborhood when comparing each pair of vertices, and match the vertices if the number of common seeds exceeds the given threshold. The reason we considered 2-hop neighborhood instead of 1-hop neighborhood was to make the algorithm work even from very sparse regime of inter-community edge density, $n_{\min} q = C(\log\log n_{\min})^2$ for $C > 0$. In the simulation, since the inter-community edge density is not very sparse, it is sufficient to count the number of seeds in the 1-hop neighborhood as the seeded matching algorithm proposed by Korula & Lattanzi (2013). Thus, we use the greedy seeded matching algorithm, where we match the pair of vertices with the maximum number of common seeds in the 1-hop neighborhood among the set of remaining vertices as we match the nodes one by one.

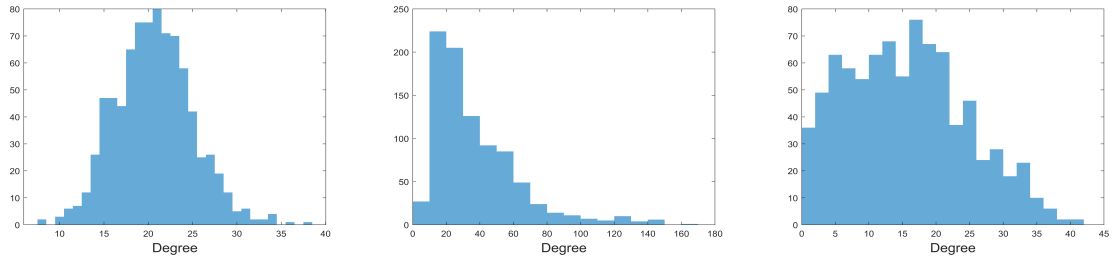
- Refinement matching : After the initial matching, refinement matching is performed on each community for T rounds. Theorem 2.4 in (Mao et al., 2021a) shows that exact matching is achievable from partial matching using refinement matching by Algorithm 4. In Algorithm 4, the refinement step runs over $\lceil \log_2 n \rceil$ rounds by checking whether the

[1] https://www.kaggle.com/datasets/ayushkalla1/rotten-tomatoes-movie-database?resource=download

[2] https://www.kaggle.com/datasets/jyoti1706/imdbmoviesdataset?datasetId=9670

number of common pairs in the 1-hop neighborhood of the matched pairs exceeds a certain threshold and whether it is below a certain threshold for all unmatched pairs. Since setting the threshold requires knowledge of the model parameters (p, q), instead we solve the linear assignment problem $\pi_t = \arg\max_{\pi_* \in \mathcal{S}_n} \sum_i |\pi_{t-1}^{-1}(\mathcal{N}_{G^\pi}(\pi_*(i))) \cap \mathcal{N}_{G'}(i)|$ where $\mathcal{S}_n$ is the permutation matrix. Appendix §A.4 shows the effect of the number of iteration rounds T on the final accuracy of the matching.

A.2. Degree Distribution of Datasets



(a) SBM (b) BlogCatalog (c) Movie

Figure 2: Degree histogram for three different datasets

Figure 2 shows the histograms of vertex degrees within the initial comparison set C_k for three different datasets, considered in Sec. 4. Unlike the 'SBM' dataset, where we sampled the network from the statistical model and thus the degree distribution within each community follows the binomial distribution, the degree distribution in real datasets deviates from the SBM. In particular, the degree distribution of the 'BlogCatalog' dataset is very different from the SBM and follows a power law. Since Algorithm 2 calculates the normalized distance between the signature vectors assuming the binomial degree distribution, we modify the definition of the signature vectors by using the logarithm of the degree instead of the degree itself for this dataset.

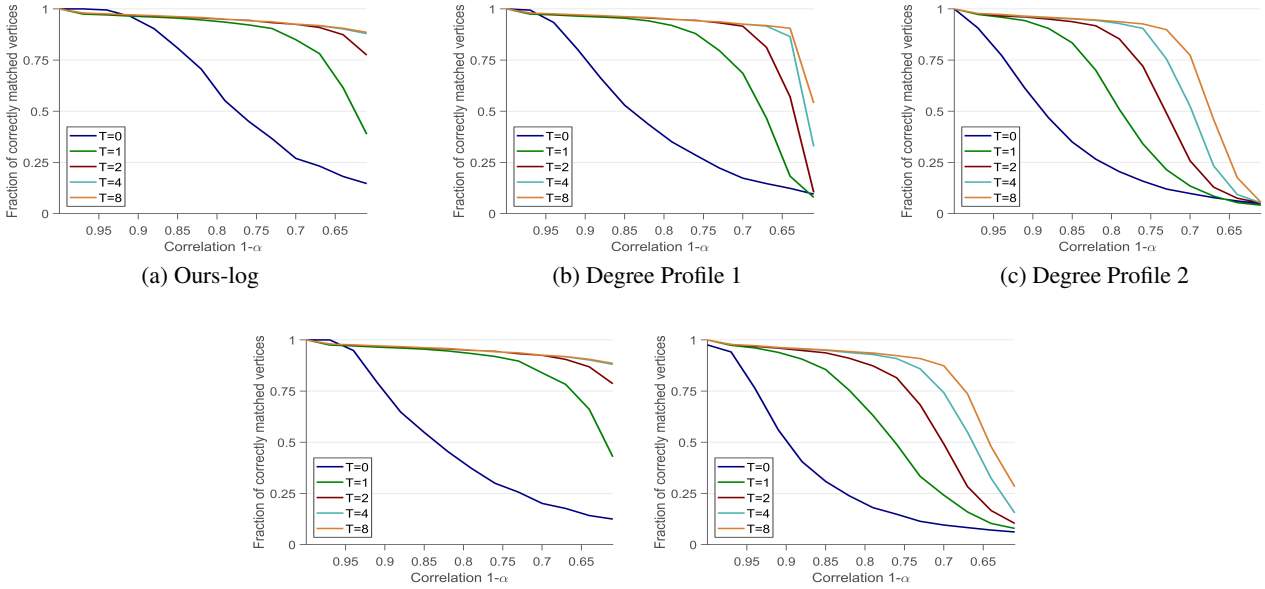
A.3. Impact of Hyperparamters k' and ℓ for Partition Tree Algorithm

Table 3: Fraction of correctly matched vertices per k' and ℓ

	SBM ($1-\alpha=0.9$)		BlogCatalog ($1-\alpha=0.85$)		Movie	
k' \ ℓ	1	2	1	2	1	2
1	0.0099	0.0148	0.0605	0.1283	0.0342	0.0565
2	0.0298	0.0683	0.1552	0.3582	0.0907	0.2167
3	0.0740	0.2339	0.2622	0.6317	0.1567	0.2945
4	0.1505	0.5598	0.3747	0.8015	0.1908	0.3557
5	0.2241	0.8252	0.4452	0.8989	0.2073	0.3416

In our partition tree algorithm with $2^{k'}$-ary tree of depth-ℓ, the two parameters (k', ℓ), which determine the dimension of the signature vectors, can be considered as hyperparameters. In Table 3, we show the impact of these parameters by reporting the performance of almost exaction matching over C_k for different (k', ℓ) pairs, where the numbers indicate the fraction of correctly matched vertices over C_k after solving the linear assignment problem for the similarity matrix obtained by the signature vectors from the partition tree. The numbers reported for the 'SBM' and 'BlogCatalog' datasets are the results averaged over 20 independent runs by randomly sampling the networks, while the numbers reported for the 'Movie' dataset are the result from a given fixed correlated networks (without any sampling or averaging). For a fixed ℓ, it can be seen that the accuracy increases with k' in most cases. For a fixed k', the accuracy increases as ℓ increases in all the cases considered. When we generate a $2^{k'}$-ary partition tree of depth ℓ, if $k'\ell > 8$, the number of leaves $2^{k'\ell}$ in a partition tree becomes larger than the number of vertices in C_k (which is about 830), and thus some of the leaf nodes become empty. So we choose $(k', \ell) = (4, 2)$ in our experiment.

A.4. Final Accuracy of Refinement Matching with Different Iterations on BlogCatalog Dataset



(a) Ours-log (b) Degree Profile 1 (c) Degree Profile 2

(d) GRAMPA 1 (e) GRAMPA 2

Figure 3: The fraction of correctly matched vertices in the final matching with respect to the correlation $1-\alpha$, for different number T of iteration rounds.

Figure 3 shows the effect of iteration rounds on the refinement matching for the BlogCatalog Dataset. In most experiments, the accuracy improves as the number of iteration rounds T increases, but each algorithm requires a different number of rounds to converge, depending on the initial accuracy of the algorithm. For example, for $T = 2$ and $0 \leq 1-\alpha \leq 0.65$, our algorithm and GRAMPA 1 converge, while others require more iterations to converge. This result implies that our algorithm requires less computation time for the final matching.

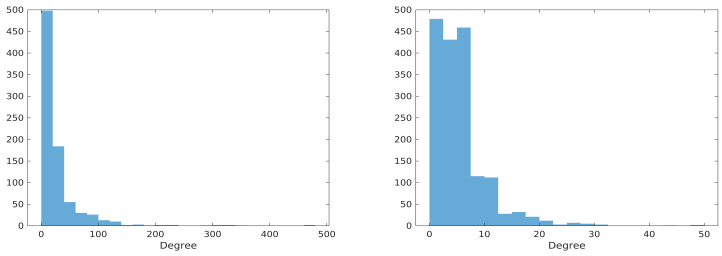
A.5. Additional Experiments

We compare our algorithm with baselines on two additional real-world datasets summarized in Table 4. For the 'Flickr' dataset (Li et al., 2015), the parent graph is given by the network of the image hosting and sharing website, where each user is treated as a vertex and the edges are connected between the vertices if they follow each other. This data consists of 7,575 vertices, 239,738 edges, and 9 communities. ACM-DBLP data (Zhang & Tong, 2019) consists of two correlated networks regarding the papers posted on ACM and DBLP in 2016, respectively. Authors are treated as vertices and an edge is added when two authors are co-authors. The community label of each author is given by their research area. This data consists of two graphs of different sizes, 9,872 / 9,916 vertices with $\sim$ 6,000 ground-truth pairs. Since the baseline algorithm GRAMPA runs on correlated graphs of the same size, we only choose vertices in the ground-truth pairs to compare the performance of our algorithm and the baselines. Thus, we use the correlated graphs consisting of 6,299 vertices, 24,822 / 27,892 edges, and 4 communities. The correlations between the two graphs are 0.9511 / 0.8464 from the perspective of each graph.

Figure 4 shows the histograms of vertex degrees within the initial comparison set C_k for these two datasets. We use the

Table 4: Details of additional datasets.

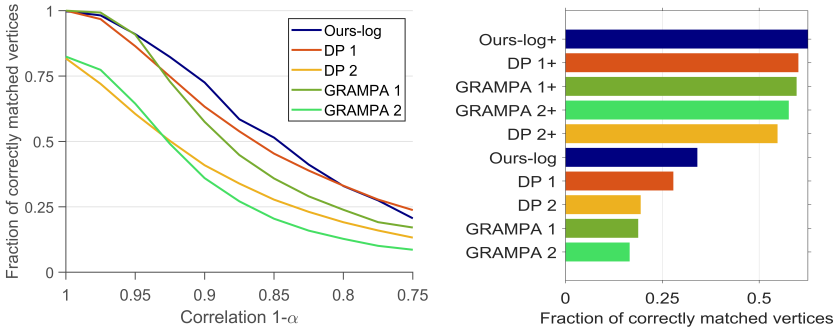
Type	Model/Source	n	# of edges	k
Sampled	Flickr	7,575	240k	9
Correlated	ACM	6,299	25k	4
	DBLP	6,299	28k	4



(a) Flickr (b) ACM-DBLP

Figure 4: Degree histogram for additional datasets

logarithm of the degree when computing the signature vectors in our algorithm, since the degree distributions of these two additional datasets follow a power law like the BlogCatalog dataset considered in the main experiment. In the case of the ACM-DBLP dataset, each community has a size of about 1,600 vertices, which is twice the size of the BlogCatalog data, and the average degree within the comparison set C_k is lower, indicating a sparser network compared to the BlogCatalog. Therefore, we set the hyperparameter $\ell = 3$ instead of $\ell = 2$. In addition, since the maximum number of communities that can be used to construct signatures is $4 - 1$ (the number of communities -1), we set $k' = 3$. On the other hand, in the case of the Flickr dataset, we set $\ell = 2$, the same as before. When comparing the degree distributions (Figure 4a and Figure 2b), we observe that the Flickr data has a long-tail distribution compared to the BlogCatalog data. In order to evenly partition the neighbors of high-degree vertices belonging to the long tail for the Flickr dataset, we set k' to be a larger value, 6 instead of 4.



(a) Flickr (b) ACM-DBLP

Figure 5: Comparison of our partition tree algorithm with other baselines, GRAMPA and Degree Profile (DP) for matching networks of (a) Flickr (b) ACM-DBLP

In Figure 5a, the result shows the fraction of correctly matched vertices within the comparison set C_k on the Flickr dataset with different correlations, averaged over 20 independent runs. Although the performance gap between our algorithm and other algorithms gets smaller, our algorithm still outperforms other algorithms on this dataset with a long-tail distribution of vertex degrees.

For ACM-DBLP data, Figure 5b shows the fraction of correctly matched vertices within the comparison set C_k (without $+$ mark) and the accuracy after final matching, which is indicated with $+$ mark. For this dataset, since there are some vertices that do not have any intra-community edges, we added the refinement matching step using an adjacency matrix for all vertices in the final step. For this dataset, our algorithm also achieves the best performance.

B. Proof of Corollary 1.2

When the community structure is not known in both graphs G^π and G', we first perform community detection on each graph separately. Then, we apply our graph matching algorithm to the two correlated graphs, using the recovered community labels. To ensure the independence between the two-step procedures, we use a random edge splitting approach in both graphs G^π and G'. Specifically, we randomly partition the edges of each graph into two sub-graphs (G_1^π, G_2^π) and (G_1', G_2') with probabilities β and $1-\beta$ respectively, where $\beta \in (0,1)$. We then apply the community detection algorithm from Yan et al. (2018) to G_1^π and G_1', and our graph matching algorithm to (G_2^π, G_2'). This ensures that the edges used for community detection are not reused in the graph matching process.

Note that both G_1^π and G_1' are graphs generated by the stochastic block model, with intra- and inter-community edge densities of $p\beta$ and $q\beta$, respectively. The result from (Yan et al., 2018) can be expressed in the following simple form.

Theorem B.1 (Theorem 1 in (Yan et al., 2018))**.** *There exists an absolute constant $C_0 > 0$ with the properties below. Let $n_{\min}$ be the minimum community size. Let B be the $k \times k$ symmetric matrix, whose (i,j)-th entry, B_{ij}, is equal to the probability that a node in community i is connected to a node in community j. If*

$$\min_k \left(B_{kk} - \max_{l \neq k} B_{kl} \right) \geq 2\sqrt{6 \log n} \max_k \sqrt{B_{kk}/n_k} + 6 \max_{1 \leq k < l \leq r} \sqrt{B_{kl} \log n / n_{\min}} + C_0 \sqrt{(n/n_{\min}^2)\,(\max B_{kk})},$$

then the exact community detection is possible with high probability by using the semidefinite programming (SDP) proposed in (Yan et al., 2018).

By applying the result from Yan et al. (2018), under the condition (1.4), we can complete the community detection on the graphs G_1^π and G_1' with high probability, where M_1 is a sufficiently large constant depending only on β. Because of the assumption that the sizes of all the communities are different, once the communities are detected, we can complete the matching between the communities of the two graphs just by comparing their sizes.

Note that $G_2^\pi = G^\pi \backslash \mathcal{E}(G_1^\pi)$ and $G_2' = G' \backslash \mathcal{E}(G_1')$ where $\mathcal{E}(G)$ is the edge set of G. Then, $(1-\alpha)(1-\beta)$ is the correlation between the two graphs G_2^π and G_2', and the community structures are now revealed in both G_2^π and G_2'. If we choose $\beta = \alpha_1/2$ and $\alpha \in (0, \alpha_1/2)$, the correlation $(1-\alpha)(1-\beta)$ is greater than $1-\alpha_1$. The conditions (1.1), (1.2), and (1.3) in Theorem 1.1 are all satisfied on the graphs G_2^π and G_2' for any constant β when n is sufficiently large. Moreover, combining the maximum density condition in (1.1) with the condition for community detection in (1.4) gives $n_{\min} = \Omega(n^{20/21})$, which is a stricter condition than $n_{\min} = \Omega(n^{10/19})$. Thus, all conditions in Theorem 1.1 are satisfied in both G_2^π and G_2', showing that the exact matching of (G_2^π, G_2') is possible by using our graph matching algorithm. This completes the proof.

C. Proof of Theorem 2.4

Let the set of seed pairs given by $\pi_s : [n_s] \to [n_s]$ over C_s be

$$S := \{(\pi_s(u_1), u_1), (\pi_s(u_2), u_2), \ldots, (\pi_s(u_{n_s}), u_{n_s})\}. \tag{C.1}$$

Let V_t and V_t' be the vertex sets in the community C_t of the graph G^π and G', respectively. Let $\mathcal{E}(G^\pi)$ and $\mathcal{E}(G')$ be the sets of edges in the graphs G^π and G', respectively. For every vertex pair $(i, i') \in V_t \times V_t'$, we define the weight function

$$w(i,i') := \left| \{ j \in [n_s] \mid \exists z \in \mathcal{N}_{G^\pi(C_t)}(i) \text{ and } \exists z' \in \mathcal{N}_{G'(C_t)}(i') \text{ such that } (z, \pi_s(u_j)) \in \mathcal{E}(G^\pi) \text{ and } (z', u_j) \in \mathcal{E}(G') \} \right|. \tag{C.2}$$

For a correct pair $(\pi(i), i) \in V_t \times V_t'$, the distribution of $\left| \mathcal{N}_{G^\pi(C_t)}(\pi(i)) \cap \mathcal{N}_{G'(C_t)}(i) \right|$ is given by Binomial$(n_t, p(1-\alpha))$. Thus, applying the bounds on binomial tail (Theorem I.1), we have

$$\left| \mathcal{N}_{G^\pi(C_t)}(\pi(i)) \cap \mathcal{N}_{G'(C_t)}(i) \right| \geq \frac{1}{2} n_t p (1-\alpha) \tag{C.3}$$

with probability at least $1 - 1/n_t^2$, where $\alpha < 1/10$ and $n_t p \geq K \log n_t$ for a sufficiently large constant $K > 0$. Thus, for any seed pair $(\pi_s(u_j), u_j)$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\exists z \in \mathcal{N}_{G^\pi(C_t)}(\pi(i)) \cap \mathcal{N}_{G'(C_t)}(i) \text{ s.t. } (z, \pi_s(u_j)) \in \mathcal{E}(G^\pi) \text{ and } (z, u_j) \in \mathcal{E}(G')\right\} \\
&= 1 - (1 - q(1-\alpha))^{\left|\mathcal{N}_{G^\pi(C_2)}(\pi(i)) \cap \mathcal{N}_{G'(C_2)}(i)\right|} \\
&\geq 1 - (1 - q(1-\alpha))^{\frac{1}{2} n_t p (1-\alpha)} \\
&= q(1-\alpha)\left(1 + (1 - q(1-\alpha)) + (1 - q(1-\alpha))^2 + \cdots + (1 - q(1-\alpha))^{\frac{1}{2} n_t p (1-\alpha) - 1}\right) \\
&\overset{(a)}{\geq} \frac{1}{2} n_t p q (1-\alpha)^2 \left(1 - \frac{1}{2} n_t p q (1-\alpha)^2\right) \overset{(b)}{\geq} \frac{1}{3} n_t p q (1-\alpha)^2 \geq \frac{1}{4} n_t p q.
\end{aligned} \tag{C.4}$$

where (a) holds since $(1-x)^n \geq 1 - nx$ for any $x < 1$, (b) holds since $n_t p q \leq 2/3$, and the last inequality holds since $\alpha < 1/10$. Since the weight function $w(\pi(i), i)$ counts the number of seed pairs connected to at least one neighboring node pair of $(\pi(i), i)$, we have

$$\mathbb{P}\left\{w(\pi(i), i) \leq \frac{1}{8} n_s n_t p q\right\} \leq \mathbb{P}\left\{\text{Bin}\left(n_s, \frac{1}{4} n_t p q\right) \leq \frac{1}{8} n_s n_t p q\right\} \overset{(a)}{\leq} \frac{1}{n_t^2}, \tag{C.5}$$

where (a) holds by Theorem I.1 with $n_s n_t p q \geq K \log n_t$ for a sufficiently large constant $K > 0$. By applying a union bound over all $i \in V_t$, we have

$$w(\pi(i), i) > \frac{1}{8} n_s n_t p q \tag{C.6}$$

for all correct pairs with probability at least $1 - \frac{2}{n_t}$.

For a wrong pair $(\pi(i), i') \in V_t \times V_t'$ with $i \neq i'$, we have $\left|\mathcal{N}_{G^\pi(C_t)}(\pi(i)) \cap \mathcal{N}_{G'(C_t)}(i')\right| \sim \text{Bin}\left(n_t, p^2(1-\alpha)^2\right)$. Thus,

$$\begin{aligned}
\mathbb{P}\left\{\left|\mathcal{N}_{G^\pi(C_t)}(\pi(i)) \cap \mathcal{N}_{G'(C_t)}(i')\right| \geq 4 n_t p^2 \vee 8 \log n_t\right\} &= \mathbb{P}\left\{\text{Bin}\left(n_t, p^2(1-\alpha)^2\right) \geq 4 n_t p^2 \vee 8 \log n_t\right\} \\
&\leq \mathbb{P}\left\{\text{Bin}\left(n_t, p^2\right) \geq 4 n_t p^2 \vee 8 \log n_t\right\} \overset{(a)}{\leq} \frac{1}{n_t^3},
\end{aligned} \tag{C.7}$$

where (a) holds from the binomial tail bounds (Theorem I.1). Moreover, from $\left|\mathcal{N}_{G_0(C_t)}(i)\right| \sim \text{Bin}\left(n_t, p/(1-\alpha)\right)$, we can show that

$$\mathbb{P}\left\{\left|\mathcal{N}_{G_0(C_t)}(i)\right| \geq 2 n_t p\right\} = \mathbb{P}\left\{\text{Bin}\left(n_t, p/(1-\alpha)\right) \geq 2 n_t p\right\} \leq \frac{1}{n_t^3}, \tag{C.8}$$

where the last inequality holds by Theorem I.1 since $\alpha < 1/10$ and $n_t p \geq K \log n_t$ for a sufficiently large constant $K > 0$. (C.8) implies that

$$\left|\mathcal{N}_{G^\pi(C_t)}(\pi(i))\right| \vee \left|\mathcal{N}_{G'(C_t)}(i')\right| \leq 2 n_t p$$

holds with probability at least $1 - \frac{1}{n_t^3}$. Thus, for any seed pair $(\pi_s(u_j), u_j))$ and $i \neq i'$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\exists z \in \mathcal{N}_{G^\pi(C_t)}(\pi(i)) \text{ and } \exists z' \in \mathcal{N}_{G'(C_t)}(i') \text{ s.t. } (z, \pi_s(u_j)) \in \mathcal{E}(G^\pi) \text{ and } (z', u_j) \in \mathcal{E}(G')\right\} \\
&\overset{(a)}{\leq} \left\{1 - (1 - q(1-\alpha))^{4 n_t p^2 \vee 8 \log n_t}\right\} + \left\{1 - (1-q)^{2 n_t p}\right\}^2 \overset{(b)}{\leq} 4 n_t p^2 q \vee 8 q \log n_t + 4 n_t^2 p^2 q^2 \\
&\leq 16\left(q \log n_t \vee n_t^2 p^2 q^2 \vee n_t p^2 q\right) \leq \frac{1}{16} n_t p q.
\end{aligned} \tag{C.9}$$

The inequality (a) holds since the first term is the probability that there exists a common neighbor pair that is connected to the seed pair and the second term is the probability that two different neighboring nodes $z \in \mathcal{N}_{G^\pi(C_t)}(\pi(i))$ and $z' \in \mathcal{N}_{G'(C_t)}(i')$ are connected to the seed pair. The inequality (b) holds since $(1-x)^n \geq 1 - nx$ for $x < 1$, and the last inequality holds since $n_t p \geq K \log n_t$ for a sufficiently large constant K, $n_t p q \leq 1/256$ and $p \leq 1/256$. Therefore, applying Theorem I.1, we get

$$\mathbb{P}\left\{w(\pi(i), i') \geq \frac{1}{8} n_s n_t p q\right\} \leq \mathbb{P}\left\{\text{Bin}\left(n_s, \frac{1}{16} n_t p q\right) \geq \frac{1}{8} n_s n_t p q\right\} \leq \frac{1}{n_t^3}, \tag{C.10}$$

where $n_s n_t pq \geq K \log n_t$ for a sufficiently large constant $K > 0$. Applying the union bound over $(\pi(i), i') \in V_t \times V_t'$ with $i \neq i'$, we have

$$w(\pi(i), i') < \frac{1}{8} n_s n_t pq \tag{C.11}$$

for all wrong pairs with probability at least $1 - \frac{2}{n_t}$. By combining (C.6) and (C.11), the proof is complete.

D. Proof of Corollary 2.5

Applying Corollary 2.2, we obtain the permutation $\hat{\pi}_k$ over C_k such that $\hat{\pi}_k = \pi|_{C_k}$ with probability at least $1 - 2n_{\min}^{-D}$. From the assumptions (1.1), (1.2) and (1.3) in Theorem 2.1, all conditions of Theorem 2.4 except $n_t pq \leq 1/256$ are satisfied. The condition $n_t pq \leq 1/256$ can be satisfied when $n_{\min} \geq M_1 n^{10/19}$ for a sufficiently large constant $M_1 > 0$, since for any $t \in [k]$, we have

$$n_t pq \leq npq \leq np^2 \overset{(a)}{\leq} M_1^{-19/10} n_{\min}^{19/10} p^2 = M_1^{-19/10} \frac{(n_{\min} p)^2}{n_{min}^{1/10}} \leq M_1^{-19/10}, \tag{D.1}$$

where the inequality (a) holds since $n_{\min} \geq M_1 n^{10/19}$ and the last inequality holds since $n_{\min} p \leq n_{\min}^{1/20}$ from (1.1).

Let us choose a community C_r that has not been used to generate the signature vector for vertices in C_k. Then, applying Theorem 2.4 with C_k, we can obtain the permutation $\hat{\pi}_r$ over C_r such that $\hat{\pi}_r = \pi|_{C_r}$ with probability at least $1 - 4/n_r$. Similarly, we can complete the exact matching over the remaining communities with probability at least $1 - \sum_{i=1}^{k} \frac{4}{n_i}$ by applying Theorem 2.4 with C_r as seeds. Since $n_{\min} \geq M_1 n^{10/19}$, the number of communities k should be less than $\frac{1}{M_1} n^{9/19}$. Thus, we have

$$1 - \sum_{i=1}^{k} \frac{4}{n_i} \geq 1 - \frac{4k}{n_{\min}} \geq 1 - \frac{1}{n^{1/19}}, \tag{D.2}$$

which completes the proof.

Remark D.1 (Relation between q and $n_{\min}$)**.** The reason we proposed a new seeded matching algorithm (Algorithm 3 beyond (Korula & Lattanzi, 2013)) and proved Theorem 2.4 was to do the seeded matching at a smaller q such as $n_{\min} q = M'(\log\log n_{\min})^2$ for a sufficiently large constant $M' > 0$. But if q is large enough such as $n_{\min} q \geq C \log n$ for a sufficiently large constant $C > 0$, then the seeded matching can be performed by simply counting the number of common seed pairs in the 1-hop neighbors as proposed by Korula & Lattanzi (2013). In this case, $n_{\min} \geq M_1 n^{10/19}$ is not required.

D.1. Balanced Community Size

In Remark 2.7, we discussed the case where the community sizes are approximately balanced, i.e., $n_{\min} = \Theta(n/k)$, and derived the conditions to guarantee the exact recovery of the underlying permutation for the correlated SBMs with constant correlation within a low-order polynomial-time complexity. Next, we prove that when the community label of each vertex is sampled from a probability distribution $[r_1, r_2, \ldots, r_k]$ with each $r_i = \Theta(1/k)$ the community sizes are approximately balanced and the sizes of all communities are different with high probability as mentioned in Remark 1.3

Lemma D.2. *Suppose that there are n vertices and k communities. Let r_i be the probability that each vertex belongs to community C_i, where*

$$\sum_{i=1}^{k} r_i = 1 \text{ and } r_i = \Theta\left(\frac{1}{k}\right). \tag{D.3}$$

If the number of communities

$$k = o(n^{1/5}) \tag{D.4}$$

then the communities are approximately balanced, i.e., $n_i = \Theta(n/k)$ for all $i \in [k]$, and the sizes of all communities are different with high probability as $n \to \infty$.

Proof. Without loss of generality, assume that $r_1 \geq \ldots \geq r_k$. Let n_i be the community size of C_i and $n_{\min}$ be the smallest community size. Define an event

$$\mathcal{F} := \left\{ n_{\min} \geq \frac{1}{2} n r_k \right\}.$$

Then, we can get

$$\begin{aligned}\mathbb{P}\{\mathcal{F}^c\} &\le \sum_{i=1}^{k} \mathbb{P}\{n_i \le \frac{1}{2}nr_k\} \le k\mathbb{P}\left\{n_k \le \frac{1}{2}nr_k\right\} \\ &\le k\mathbb{P}\left\{\mathrm{Bin}(n, r_k) \le \frac{1}{2}nr_k\right\} \overset{(a)}{\le} k\exp(-3\log n) \le \frac{k}{n^3}.\end{aligned} \tag{D.5}$$

The inequality (a) holds by Theorem I.1 with $\frac{n}{k} = \omega(n^{4/5})$. Moreover, we can show that $n_{\max} \le 2nr_1$ with probability at least $1 - \frac{k}{n^3}$ using a similar approach as in (D.5). Therefore, $n_i = \Theta(n/k)$ holds with high probability as $n \to \infty$.

For any $i, j \in [k]$, $i \ne j$, we have

$$\begin{aligned}\mathbb{P}\{n_i = n_j|\mathcal{F}\} &\le \sum_{t \ge \frac{1}{2}nr_k} \mathbb{P}\{n_i = n_j | n_i + n_j = 2t\}\mathbb{P}\{n_i + n_j = 2t\} \\ &\le \max_{t \ge \frac{1}{2}nr_k} \mathbb{P}\{n_i = n_j | n_i + n_j = 2t\} \\ &\le \max_{t \ge \frac{1}{2}nr_k} \binom{2t}{t}\left(\frac{r_i}{r_i + r_j}\right)^t \left(\frac{r_j}{r_i + r_j}\right)^t \\ &\le \max_{t \ge \frac{1}{2}nr_k} \binom{2t}{t}\frac{1}{2^{2t}} \overset{(a)}{=} O\left(\frac{1}{\sqrt{nr_k}}\right).\end{aligned} \tag{D.6}$$

The equality (a) holds by Stirling's approximation. Thus, we have

$$\mathbb{P}\{n_i = n_j\} \le \mathbb{P}\{n_i = n_j|\mathcal{F}\} + \mathbb{P}\{\mathcal{F}^c\} = O\left(1/\sqrt{nr_k}\right) = O\left(\sqrt{k/n}\right). \tag{D.7}$$

Applying a union bound over i, j, we can show that the community sizes are all different with probability $1 - O\left(k^2\sqrt{\frac{k}{n}}\right)$. We have $k^2\sqrt{\frac{k}{n}} \to 0$ as $n \to \infty$ since (D.4). □

In Remark 2.7, we assumed that $k \le C'\sqrt{\frac{n}{p}}(p - q)$. On the condition $np \le n^{1/20}$, we have

$$C'\sqrt{\frac{n}{p}}(p - q) \le C'\sqrt{np} \le n^{1/40} = o(n^{1/5})$$

as $n \to \infty$. Therefore, the condition (D.4) in Lemma D.2 can be satisfied.

E. Proof of Theorem 2.1

From now on, assume that the permutation $\pi : [n] \to [n]$ is the identity, so that $G^\pi = G$. Let C_k be the smallest community and m denote the size of the community C_k, i.e, $m = n_{\min}$.

In proving Theorem 2.1 (almost exact matching of $G^\pi(C_k)$ and $G'(C_k)$), we follow the proof structure mostly similar to that of Mao et al. (2021a), where the almost exact matching of the correlated ER model was proved based on the correlation between the signature vectors of the correct pair of vertices, where the signature vector is defined in terms of the binary partition tree rooted from each vertex. The main difference of our analysis from that of Mao et al. (2021a) is that we use both inter- and intra-community edges when constructing the $2^{k'}$-ary partition tree, while that of Mao et al. (2021a) uses only inter-community edges (since there is no community structure in the ER model) both in generating the binary tree and in partitioning the vertices within the tree. In our analysis, the intra-community edges are used to generate the tree structure, while the inter-community edges are used to partition the vertices at each level of the tree into different nodes based on their inter-community edge degrees. The proof becomes simpler for our analysis of the correlated SBMs due to this difference.

In this section, we present two main lemmas, Lemma E.1 and Lemma E.2, to prove Theorem 2.1. Lemma E.1 shows that for most correct vertex pairs, the value of the normalized distance $\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i)\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i)}$ between the signature vectors in Algorithm 2 is less than the threshold $2w\left(1 - \frac{1}{\sqrt{\log m}}\right)$ with high probability.

Lemma E.1. *For any constants $C, D > 0$, there exist constants $R, Q_1, Q_2, m_0, \alpha_1, c > 0$ with the properties below. Consider the graphs G^π and G', which are generated from the correlated Stochastic Block Models (SBMs) described in Section 1.1 with correlation $1-\alpha$. Consider a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$, where w is an integer. Suppose that $m \geq m_0$, $\alpha \in (0, \alpha_1)$, $w \geq (\log m)^4$, and the following conditions hold:*

$$(\log m)^{1.1} \leq mp \leq m^{1/20},$$
$$mq \geq Q_1 k'^2 \ell^2,$$
$$2\log\left(w^4(\log m)\right) \leq k'\ell \leq C\log\log m, \quad k' \leq Q_2 \log mp, \quad \ell \leq \frac{\log m}{R\log mp} \wedge C\log\log m.$$

Then, there are at least $m - m^{1-c}$ vertices $i \in C_k$ satisfying

$$\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i)\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i)} \leq 2w\left(1 - \frac{1}{(\log m)^{0.1}}\right), \tag{E.1}$$

with probability at least $1 - m^{-D}$.

The proof of Lemma E.1 can be found in Section G.

Lemma E.2, on the other hand, demonstrates that for most of the incorrect pairs (i, i') in C_k, the normalized distance value $\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')}$ between the signature vectors exceeds the threshold $2w\left(1 - \frac{1}{\sqrt{\log m}}\right)$ with high probability.

Lemma E.2. *For any constants $C, D > 0$, there exist constants $R, Q_1, Q_2, m_0, \alpha_1, c > 0$ with the properties below. Consider the graphs G^π and G', which are generated from the correlated Stochastic Block Models (SBMs) described in Section 1.1 with correlation $1-\alpha$. Consider a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$, where w is an integer. Suppose that $m \geq m_0$, $\alpha \in (0, \alpha_1)$, $w \geq \lfloor (\log m)^5 \rfloor$, and the following conditions hold:*

$$(\log m)^{1.1} \leq mp \leq m^{1/20},$$
$$mq \geq Q_1 k'^2 \ell^2,$$
$$k'\ell \leq C\log\log m, \quad k' \leq Q_2 \log mp, \quad \ell \leq \frac{\log m}{R\log mp} \wedge C\log\log m.$$

Then, there exists a subset $\mathcal{I} \subset C_k$ with a size $|\mathcal{I}| \geq m - m^{1-c}$ such that for any $i, i' \in \mathcal{I}$, $i \neq i'$, we have

$$\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')} \geq 2w\left(1 - \frac{1}{(\log m)^{0.9}}\right), \tag{E.2}$$

with probability at least $1 - m^{-D}$.

The proof of Lemma E.2 can be found in Section H.

Proof of Theorem 2.1. We will set the order of ℓ as large as possible to minimize the order of k' (or k). Let $\ell = \left\lceil \frac{\log m}{40\log mp} \right\rceil \wedge \lceil 42\log\log m \rceil$ and $k' = \left\lceil 1680\log\log m \frac{\log mp}{\log m} \right\rceil$. We consider a random subset $J \in \{-1,1\}^{k'\ell}$ of a size $|J| = 2w$ where $w = \lfloor (\log m)^5 \rfloor$. By Lemma E.1 and E.2, we can find a subset $\mathcal{I} \subset C_k$ with its size $|\mathcal{I}| \geq m - m^{1-c}$ for a positive constant c such that $\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i)\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i)} < 2w\left(1 - \frac{1}{\sqrt{\log m}}\right)$ and $\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')} > 2w\left(1 - \frac{1}{\sqrt{\log m}}\right)$ for any distinct $i, i' \in \mathcal{I}$. Hence, by running Algorithm 2, we can get a matrix B satisfying that

$$B_{\pi(i'),i} = \begin{cases} 1 & \text{if } i = i' \in \mathcal{I}, \\ 0 & \text{if } i, i'(i \neq i') \in \mathcal{I}. \end{cases}$$

By Proposition 2.2 in (Mao et al., 2021a) with matrix B as an input, we can obtain a permutation $\hat{\pi}$ on community C_k that satisfies

$$|i \in [m] : \hat{\pi}(i) \neq \pi(i)| \leq 4m^{1-c}. \tag{E.3}$$

This completes the proof. □

F. Structural Properties of Partition Tree

We analyze the statistical behavior of large neighborhoods of vertices in Erdős-Rényi (ER) graph and the partition trees rooted from each vertex, which will be used to prove the main theoretical results. Section F.1 is dedicated to the statement of lemmas concerning the statistics of the large neighborhoods in the ER graph $\mathcal{G}(n,p)$. In Section F.2, we show that the sizes of the nodes $T^d_{\boldsymbol{s}^d}(i)$ at each level $d \in [\ell]$ of the partition tree are well balanced for every $\boldsymbol{s}^d \in \{-1,1\}^{k'd}$. Furthermore, in Section F.3 we establish that the ℓ-neighborhoods of a majority of vertices in C_k form a tree with high probability.

F.1. Large Neighborhoods in ER Graph

In this subsection, we state lemmas that analyze the statistics of large neighborhoods in Erdős-Rényi (ER) graph $\mathcal{G}(n,p)$.

Lemma F.1 (Sizes of neighbors)**.** *For any constant $D, R, \delta > 0$, there exists a constant n_0, which depends only on D, R and δ, with the properties below. Consider an Erdős-Rényi (ER) graph $G \sim \mathcal{G}(n,p)$. Assume that*

$$n \geq n_0, \quad \text{and} \quad np \geq (\log n)^{1+\delta}. \tag{F.1}$$

Then, we have that

$$|\mathcal{S}_G(i,1) - np| \leq \frac{np}{R \log\log n} \quad \forall i \in [n], \tag{F.2}$$

with probability at least $1 - n^{-D}$.

Proof. Applying Bernstein's inequality (stated in Lemma I.5), for any vertex i we have

$$\mathbb{P}\left\{|\mathcal{S}_G(i,1) - np| \geq \frac{np}{R\log\log n}\right\} \leq 2\exp\left(-R'\frac{np}{(\log\log n)^2}\right) \leq n^{-D-1}, \tag{F.3}$$

where R' depends only on R. The last inequality holds since $(\log n)^{1+\delta} \leq np$ and $n \geq n_0(D,R,\delta)$. By applying a union bound over all vertices $i \in [n]$, we can complete the proof. □

Lemma F.2 (Sizes of intersections of neighbors (Lemma 4.1 in (Mao et al., 2021a)))**.** *For any constant $D > 1$, there exist constants $K > 0$ and n_0, which depend only on D, with the properties below. Consider an Erdős-Rényi (ER) graph $G \sim \mathcal{G}(n,p)$ with $n \geq n_0$ and $np \geq \log n$. With probability at least $1 - n^{-D}$, the number of vertices in the ball of radius l centered at i is bounded by*

$$|\mathcal{B}_G(i,l)| \leq K(np)^l \quad \text{for any } i, l \in [n]. \tag{F.4}$$

Furthermore, under the event defined by (F.4), consider any positive integer m and $i, j \in [n]$ satisfying that $i \neq j$ and $G(\mathcal{B}_G(i,3m))$ forms a tree. If the distance $d = \mathrm{dist}_G(i,j) \leq 2m$, then we have

$$|\mathcal{B}_G(i,m) \cap \mathcal{B}_G(j,m)| \leq K(np)^{m-\lceil d/2 \rceil}.$$

F.2. Sizes of Nodes in the Partition Tree

We next provide the bounds on the sizes of the nodes $\{T^d_{\boldsymbol{s}^d}(i)\}_{\boldsymbol{s}^d \in \{-1,1\}^{k'd}, d \in [\ell]}$ in the $2^{k'}$-ary partition tree.

Lemma F.3 (Sizes of nodes in the partition tree)**.** *For any constants $D, R, \delta > 0$, there exist a constant m_0 depending on D, R, δ, a constant k_0 depending on δ and an absolute constant $Q > 0$ with the properties below. Let $G \sim SBM(n,p,q,k)$ with communities $(C_1, \ldots, C_k)$. Assume that the community labels $(C_1, C_2, \ldots, C_k)$ are given. Assume that*

$$m \geq m_0, \quad (\log m)^{1+\delta} \leq mp, \quad k' \leq k_0 \log mp, \quad mq \geq Qk'^2\ell^2, \tag{F.5}$$

for a fixed positive integer $\ell \leq \left\lfloor \frac{\log m}{\log mp} \right\rfloor \wedge R\log\log m$. For any $d \in [\ell]$, $\boldsymbol{s}^d \in \{-1,1\}^{k'd}$, and $i \in C_k$, recall $T^d_{\boldsymbol{s}^d}(i)$ defined in Sec. 2.2. Then, for any $i \in C_k$ satisfying that $G\left(\mathcal{B}_{G(C_k)}(i,\ell)\right)$ forms a tree, we have

$$\left|T^d_{\boldsymbol{s}^d}(i)\right| \leq 6\left(\frac{mp}{2^{k'}}\right)^d \tag{F.6}$$

with probability at least $1 - m^{-D}$.

Proof. For notational simplicity, we will use G_k instead of $G(C_k)$. We will prove (F.6) using mathematical induction. We start with the base case $d = 1$, and by applying Lemma F.1, we obtain the following result:

$$\mathbb{P}\left\{|\mathcal{S}_{G_k}(i,1)| \le mp\left(1+\frac{1}{5\ell}\right)\right\} \ge 1 - m^{-D-2} \tag{F.7}$$

since $mp \ge (\log m)^{1+\delta}$ and $m \ge m_0(D, R, \delta)$. Let $\boldsymbol{s}_1 \in \{-1,1\}^{k'}$. By applying the bounds from Lemma I.2 and I.3 for a binomial random variable, we can derive the following inequality for $j \in \mathcal{S}_{G_k}(i,1)$ and $a \in [k']$:

$$\left|\mathbb{P}\{\mathbf{Sign}(\deg_G^a(j) - n_a q) = 1\} - \frac{1}{2}\right| \le \frac{C}{\sqrt{n_a q}} \le \frac{C}{\sqrt{mq}} \tag{F.8}$$

where C is a universal constant. Therefore, we can show that

$$\left|\mathbb{P}\{\mathbf{Sign}(\deg_G^a(j) - n_a q) = \boldsymbol{s}_1(a) \text{ for all } a \in [k']\} - \frac{1}{2^{k'}}\right| \le \frac{4Ck'}{2^{k'}\sqrt{mq}} \tag{F.9}$$

by applying Lemma I.10 with the condition $mq \ge 4C^2k'^2$. If $mq \ge Qk'^2\ell^2$ for a sufficiently large constant Q, we obtain

$$\mathbb{P}\{\mathbf{Sign}(\deg_G^a(j) - n_a q) = \boldsymbol{s}_1(a) \text{ for all } a \in [k']\} \le \frac{1}{2^{k'}}\left(1+\frac{1}{5\ell}\right). \tag{F.10}$$

Thus, by applying Hoeffding's inequality (Lemma I.4), we have

$$\mathbb{P}\left\{\left|T^1_{\boldsymbol{s}_1}(i)\right| \le 2\frac{mp}{2^{k'}}\left(1+\frac{1}{\ell}\right)\right\} \overset{(a)}{\ge} 1 - \exp\left(-C'\frac{mp}{2^{2k'}}\right) \ge 1 - m^{-D-1}, \tag{F.11}$$

where the inequality (a) holds for a constant $C' > 0$ by (F.7) and (F.10). The last inequality holds since $k' \le k_0 \log mp$ for a sufficiently small constant k_0 depending on δ such that $2^{2k_0 \log mp} \le (mp)^{0.5\delta/(1+\delta)}$, $mp \ge (\log m)^{1+\delta}$ and $m \ge m_0(D, \delta)$.

Define an event

$$\mathcal{F}_d := \left\{\deg_{G_k}(j) \le mp\left(1+\frac{1}{5\ell}\right),\ \forall j \in \mathcal{B}_{G_k}(i,d)\right\}. \tag{F.12}$$

Let $\mathbb{P}_d$ denote the conditional probability given the subgraph $G_k(\mathcal{B}(i,d))$ under the condition that $\mathcal{F}_{d-1}$ holds. For $\boldsymbol{s}_u \in \{-1,1\}^{k'}$, let $\boldsymbol{s}^d = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d)$. Define

$$Q_d := 2\left(1+\frac{1}{\ell}\right)^d. \tag{F.13}$$

By Lemma F.1, we have

$$\mathbb{P}_d\left\{\mathcal{F}_d \text{ occurs} \;\middle|\; \left|T^d_{\boldsymbol{s}^d}(i)\right| \le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\} \ge 1 - m^{-D-2} \tag{F.14}$$

and

$$\mathbb{P}_d\left\{\left|\mathcal{N}_{G_k}(T^d_{\boldsymbol{s}^d}(i)) \cap \mathcal{S}_{G_k}(i,d+1)\right| \le |T^d_{\boldsymbol{s}^d}(i)| mp\left(1+\frac{1}{5\ell}\right) \;\middle|\; \left|T^d_{\boldsymbol{s}^d}(i)\right| \le Q_d\left(\frac{mp}{2^{k'}}\right)^d \text{ and } \mathcal{F}_d \text{ occurs}\right\} \ge 1 - m^{-D-2}. \tag{F.15}$$

Thus, we can get

$$\begin{aligned}&\mathbb{P}_d\left\{\left|\mathcal{N}_{G_k}(T^d_{\boldsymbol{s}^d}(i)) \cap \mathcal{S}_{G_k}(i,d+1)\right| \le 2^{k'}Q_d\left(\frac{mp}{2^{k'}}\right)^{d+1}\left(1+\frac{1}{5\ell}\right) \text{ and } \mathcal{F}_d \text{ occurs} \;\middle|\; \left|T^d_{\boldsymbol{s}^d}(i)\right| \le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\} \\ &\ge 1 - 2m^{-D-2}.\end{aligned} \tag{F.16}$$

Let $H := \mathcal{N}_{G_k}(T^d_{\boldsymbol{s}^d}(i)) \cap \mathcal{S}_{G_k}(i,d+1)$. For $\boldsymbol{s}_{d+1} \in \{-1,1\}^{k'}$ and any $j \in H$,

$$\mathbb{P}_d\{\mathbf{Sign}(\deg_G^a(j) - n_a q) = \boldsymbol{s}_{d+1}(a) \text{ for all } a \in [k']\} \le \frac{1}{2^{k'}}\left(1+\frac{1}{5\ell}\right). \tag{F.17}$$

The above result follows from (F.10). Similar to (F.11), by applying Hoeffding's inequality we have

$$\begin{aligned}&\mathbb{P}_d\left\{|T^{d+1}_{\boldsymbol{s}^{d+1}}(i)|\le Q_{d+1}\left(\frac{mp}{2^{k'}}\right)^{d+1}\middle||H|\le 2^{k'}Q_d\left(\frac{mp}{2^{k'}}\right)^{d+1}\left(1+\frac{1}{5\ell}\right),\mathcal{F}_d\text{ occurs and }|T^d_{\boldsymbol{s}^d}(i)|\le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\}\\&=\mathbb{P}_d\left\{|T^{d+1}_{\boldsymbol{s}^{d+1}}(i)|\le Q_{d+1}\left(\frac{mp}{2^{k'}}\right)^{d+1}\middle||H|\le 2^{k'}Q_d\left(\frac{mp}{2^{k'}}\right)^{d+1}\left(1+\frac{1}{5\ell}\right)\right\}\ge 1-m^{-D-2}.\end{aligned}\tag{F.18}$$

By (F.16) and (F.18), we finally get

$$\begin{aligned}&\mathbb{P}_d\left\{|T^{d+1}_{\boldsymbol{s}^{d+1}}(i)|\le Q_{d+1}\left(\frac{mp}{2^{k'}}\right)^{d+1}\text{ and }\mathcal{F}_d\text{ occurs}\middle||T^d_{\boldsymbol{s}^d}(i)|\le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\}\\&=\mathbb{P}_d\left\{|T^{d+1}_{\boldsymbol{s}^{d+1}}(i)|\le Q_{d+1}\left(\frac{mp}{2^{k'}}\right)^{d+1}\middle||H|\le 2^{k'}Q_d\left(\frac{mp}{2^{k'}}\right)^{d+1}\left(1+\frac{1}{5\ell}\right),\mathcal{F}_d\text{ occurs and }|T^d_{\boldsymbol{s}^d}(i)|\le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\}\\&\quad\times\mathbb{P}_d\left\{|H|\le 2^{k'}Q_d\left(\frac{mp}{2^{k'}}\right)^{d+1}\left(1+\frac{1}{5\ell}\right)\text{ and }\mathcal{F}_d\text{ occurs}\middle||T^d_{\boldsymbol{s}^d}(i)|\le Q_d\left(\frac{mp}{2^{k'}}\right)^d\right\}\\&\ge 1-m^{-D-1}.\end{aligned}\tag{F.19}$$

Applying a union bound over (F.11) and (F.19) for $d\in[\ell-1]$ and $\boldsymbol{s}^d\in\{-1,1\}^{k'd}$, we can get

$$\mathbb{P}\left\{|T^d_{\boldsymbol{s}^d}(i)|\le Q_d\left(\frac{mp}{2^{k'}}\right)^d\ \forall d\in[\ell],\ \boldsymbol{s}^d\in\{-1,1\}^{k'd}\right\}\ge 1-\Sigma_{d=1}^{\ell}2^{k'd}m^{-D-1}\ge 1-m^{-D},\tag{F.20}$$

where the last inequality holds since $k'\le k_0\log mp$ and $\ell\le\left\lfloor\frac{\log m}{\log mp}\right\rfloor$. Moreover, we have $Q_\ell=2\left(1+\frac{1}{\ell}\right)^\ell\le 2e\le 6$. Thus, the proof is complete. □

F.3. Tree Structure and Typical Vertices

Consider the parent graph G_0, which follows an SBM$(n,p/(1-\alpha),q/(1-\alpha),k)$ with community labels $(C_1,C_2,\dots,C_k)$. Recall that C_k is the smallest community with size m.

Following Definition 4.10 in (Mao et al., 2021a), we introduce the concept of a "typical" vertex in the parent graph, which satisfies certain conditions that will be useful for analyzing the signature vectors. Compared to Definition 4.10 in (Mao et al., 2021a), our definition has fewer conditions, due to the main difference of our algorithm that the edges used to construct the tree structure are not reused to partition the vertices at each level of the tree.

Definition F.4. We define a vertex $i\in C_k$ of the parent graph G_0 as typical with parameters $\ell\in\mathbb{N}$ and $\epsilon>0$, denoted as $i\in\mathrm{Typ}_{G_0(C_k)}(\ell,\epsilon)$, if it satisfies the following conditions:

$(A1)$ $G_0\left(\mathcal{B}_{G_0(C_k)}(i,\ell)\right)$ forms a tree.

$(A2)$ $\deg_{G_0(C_k)}(j)\le 2m\dfrac{p}{1-\alpha}$ for any $j\in\mathcal{B}_{G_0(C_k)}(i,\ell-1)$.

$(A3)$ $\deg_{G_0(C_k)}(j)>(1-\epsilon)m\dfrac{p}{1-\alpha}$ for any $j\in\mathcal{B}_{G_0(C_k)}(i,\ell-1)$.

Lemma F.5 (Lemma 4.9 in (Mao et al., 2021a))**.** *Consider an Erdős-Rényi graph $\Gamma\sim\mathcal{G}(n,r)$ for a positive integer n and $r\in(0,1)$. For any positive integer x, the probability that there exist at least $(5x)^3(\log n)^6(nr)^{3x}$ vertices $i\in\Gamma$ such that $\Gamma\left(\mathcal{B}_\Gamma(i,x)\right)$ is not a tree, is upper bounded by $\exp\left(-\log^2 n\right)$.*

The following result shows that the majority of vertices in $G_0(C_k)$ are typical, satisfying Definition F.4.

Proposition F.6. *For any constants $D,\delta,\epsilon>0$, there exist $R>1$, $c\in(0,1/2)$ and m_0 depending on D,δ,ϵ with the properties below. If*

$$m\ge m_0,\quad(\log m)^{1+\delta}\le m\frac{p}{1-\alpha}\le m^{1/20},\quad\ell\le\frac{\log m}{R\log mp},\tag{F.21}$$

then

$$\mathbb{P}\left\{\left|\mathrm{Typ}_{G_0(C_k)}(\ell,\epsilon)\right|\ge m-m^{1-c}\right\}\ge 1-m^{-D}.$$

Proof. Recall that $G_0(C_k)$ is a $\mathcal{G}(m, p/(1-\alpha))$ Erdős-Rényi graph. Thus, by applying Lemma F.5 for $x = \ell \le \frac{\log m}{R \log mp}$ with a sufficiently large R, we can obtain that there are at most

$$(5\ell)^3 (\log m)^6 \left(\frac{mp}{1-\alpha}\right)^{3\ell} \le \sqrt{m} \tag{F.22}$$

vertices whose ℓ-neighborhoods are not trees in $G_0(C_k)$ with probability at least $1 - \exp\left(-(\log m)^2\right) \ge 1 - m^{-D-1}$ for $m \ge m_0(D)$. Thus, the condition $(A1)$ holds. By Lemma F.1, it can be easily checked that the two conditions $(A2)$ and $(A3)$ are satisfied for all $i \in C_k$ with probability $1 - m^{-D-1}$ for $m \ge m_0(D, \delta, \epsilon)$. □

G. Proof of Lemma E.1

From now on, for simpleness in the notation we will use $T^d_{\boldsymbol{s}^d}(i)$ and $T'^d_{\boldsymbol{s}^d}(i)$ instead of $T^d_{\boldsymbol{s}^d}(i, G)$ and $T^d_{\boldsymbol{s}^d}(i, G')$, respectively. Consider a random subset J uniformly drawn from $\{-1, 1\}^{k'\ell}$ with cardinality $2w$, where w is a positive integer satisfying $w > 2(\log m)^2$. We say that the event $\mathcal{G}$ holds if and only if there exists a subset $\tilde{J}(i) \subset J$ such that

$$|J \backslash \tilde{J}(i)| < (\log m)^2 \tag{G.1}$$

and

$$|R_{\boldsymbol{s}^\ell}(i)| \vee |R'_{\boldsymbol{s}^\ell}(i)| \le 96 e w^4 \frac{(mp)^\ell}{4^{k'\ell}}, \quad \forall \boldsymbol{s}^\ell \in \tilde{J}(i), \tag{G.2}$$

$$|R_{\boldsymbol{s}^\ell}(i)| \vee |R'_{\boldsymbol{s}^\ell}(i)| \le 6 \frac{(mp)^\ell}{2^{k'\ell}}, \quad \forall \boldsymbol{s}^\ell \in J, \tag{G.3}$$

where

$$R_{\boldsymbol{s}^\ell}(i) := T^\ell_{\boldsymbol{s}^\ell}(i) \cap T'^\ell_{J \backslash \{\boldsymbol{s}^\ell\}}(i), \tag{G.4}$$

$$R'_{\boldsymbol{s}^\ell}(i) := T'^\ell_{\boldsymbol{s}^\ell}(i) \cap T^\ell_{J \backslash \{\boldsymbol{s}^\ell\}}(i). \tag{G.5}$$

Let us define the five conditions $(B1) - (B5)$ as follows.

$(B1)$ $G_0\left(\mathcal{B}_{G_0(C_k)}(i, \ell)\right)$ forms a tree.

$(B2)$ $\left|\mathcal{B}_{G_0(C_k)}(i, \ell)\right| \le m^{0.1}$.

$(B3)$ $\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right| \vee \left|T'^\ell_{\boldsymbol{s}^\ell}(i)\right| \le 6(mp/2^{k'})^\ell$ for all $\boldsymbol{s}^\ell \in \{-1, 1\}^{k'\ell}$.

$(B4)$ Event $\mathcal{G}$ holds.

$(B5)$ $\left|T^\ell_{\boldsymbol{s}^\ell}(i) \cap T'^\ell_{\boldsymbol{s}^\ell}(i)\right| \ge (mp/2^{k'})^\ell (1 - 6\epsilon)^{k'\ell}$ for a constant ϵ and for all $\boldsymbol{s}^\ell \in \{-1, 1\}^{k'\ell}$.

We will show that there are enough number of vertices $i \in C_k$ satisfying the conditions $(B1) - (B5)$, and for all such vertices (E.1) holds with high probability. The conditions $(B1) - (B3)$ will be shown to be held with high probability by using the analysis from Sec. F, and the conditions $(B4) - (B5)$ will be proved in Sec. G.3 and Sec. G.4, respectively.

G.1. Proof of Lemma E.1

We consider a fixed vertex $i \in C_k$, and subsets J and $\tilde{J}(i) \subset J$. Additionally, we condition on the neighborhoods $G_0(\mathcal{B}_{G_0(C_k)}(i, \ell))$, $G(\mathcal{B}_{G_0(C_k)}(i, \ell))$, and $G'(\mathcal{B}_{G_0(C_k)}(i, \ell))$ such that conditions $(B1) - (B5)$ hold. Let us define

$$\tilde{m} := m - \left|\mathcal{B}_{G_0(C_k)}(i, \ell)\right|. \tag{G.6}$$

Note that on the conditions $(B1)$ and $(B2)$, we can see that for any $j \in \mathcal{S}_{G_0(C_k)}(i, \ell)$, $\deg_{G(C_k)}(j) - 1$ follows a binomial distribution with parameters $\tilde{m}$ defined in (G.6) with $\tilde{m} \ge m - m^{0.1}$ and p. Furthermore, it is independent for each $j \in \mathcal{S}_{G_0(C_k)}(i, \ell)$. The same statement holds for $\deg_{G'(C_k)}(j) - 1$.

Lemma G.1 (Similar to Lemma 5.7 in (Mao et al., 2021a)). *For any constant $D > 0$, there exists a constant $K > 0$ which depends on D with the properties below. For a fixed vertex $i \in C_k$, subsets J and $\tilde{J}(i) \subset J$, define $R_{\boldsymbol{s}^\ell}(i)$ and $R'_{\boldsymbol{s}^\ell}(i)$ as in*

(G.4) *and* (G.5)*. Suppose that conditions* $(B1)-(B5)$ *hold. Then, we have*

$$|\eta_{\boldsymbol{s}^\ell}(i)| \vee |\eta'_{\boldsymbol{s}^\ell}(i)| \le K\frac{(mp)^{(\ell+1)/2}}{2^{k'\ell}}w^2\sqrt{\log m}, \quad \forall \boldsymbol{s}^\ell \in \tilde{J}(i) \tag{G.7}$$

$$|\eta_{\boldsymbol{s}^\ell}(i)| \vee |\eta'_{\boldsymbol{s}^\ell}(i)| \le K\frac{(mp)^{(\ell+1)/2}}{2^{k'\ell/2}}\sqrt{\log m}, \quad \forall \boldsymbol{s}^\ell \in J \tag{G.8}$$

where

$$\eta_{\boldsymbol{s}^\ell}(i) := \sum_{j\in R_{\boldsymbol{s}^\ell}(i)}\left(\deg_{G(C_k)}(j)-1-mp\right) \quad \textit{and} \quad \eta'_{\boldsymbol{s}^\ell}(i) := \sum_{j\in R'_{\boldsymbol{s}^\ell}(i)}\left(\deg_{G'(C_k)}(j)-1-mp\right) \tag{G.9}$$

with probability at least $1-m^{-D}$*.*

Proof. Since $R_{\boldsymbol{s}^\ell}(i) \subset \mathcal{S}_{G_0(C_k)}(i,\ell)$, for any $j \in R_{\boldsymbol{s}^\ell}(i)$, $(\deg_{G(C_k)}(j)-1)$ is distributed as $\mathrm{Bin}(\tilde{m},p)$ with $\tilde{m} := m - \left|\mathcal{B}_{G_0(C_k)}(i,\ell)\right| \ge m - m^{0.1}$. In addition, it is independent for each $j \in R_{\boldsymbol{s}^\ell}(i)$. Therefore we have $\sum_{j\in R_{\boldsymbol{s}^\ell}(i)}\left(\deg_{G(C_k)}(j)-1\right) \sim \mathrm{Bin}\left(\tilde{m}\left|R_{\boldsymbol{s}^\ell}(i)\right|,p\right)$. From this result, we can obtain

$$\begin{aligned}
|\eta_{\boldsymbol{s}^\ell}(i)| &= \left|\sum_{j\in R_{\boldsymbol{s}^\ell}(i)}\left(\deg_{G(C_k)}(j)-1-mp\right)\right| \\
&= \left|\sum_{j\in R_{\boldsymbol{s}^\ell}(i)}\left(\deg_{G(C_k)}(j)-1\right) - \tilde{m}p|R_{\boldsymbol{s}^\ell}(i)| - (m-\tilde{m})p|R_{\boldsymbol{s}^\ell}(i)|\right| \\
&\le K_1\left(\sqrt{mp\left|R_{\boldsymbol{s}^\ell}(i)\right|\log m}+\log m\right) + (m-\tilde{m})p\left|R_{\boldsymbol{s}^\ell}(i)\right| \\
&\le K_2\sqrt{mp\left(\left|R_{\boldsymbol{s}^\ell}(i)\right|+1\right)\log m}
\end{aligned} \tag{G.10}$$

with probability at least $1-m^{-D-1}$, where K_1 and K_2 are sufficiently large constants depending on D. In the same way, a bound for $\eta'_{\boldsymbol{s}^\ell}(i)$ can be obtained. By combining (G.10) with $(B4)$, we obtain (G.8) and (G.7). By applying a union bound over $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$, we can conclude the proof. □

For a fixed vertex i, we omit (i) from $T_{\boldsymbol{s}^\ell}(i)$, $R_{\boldsymbol{s}^\ell}(i)$, $f_{\boldsymbol{s}^\ell}(i)$, and $\mathrm{v}_{\boldsymbol{s}^\ell}(i)$ for the sake of brevity in the notation.

Lemma G.2 (Similar to Lemma 5.9 in (Mao et al., 2021a))**.** *Assume a realization of edges between* $(R_{\boldsymbol{s}^\ell} \cup R'_{\boldsymbol{s}^\ell})$ *and* $\mathcal{S}_{G_0(C_k)}(i,\ell+1)$ *in the graphs* G_0, G*, and* G'*. Consider the corresponding conditional probability and expectation, denoted by* $\hat{\mathbb{P}}$ *and* $\hat{\mathbb{E}}$*, respectively. Then, for any* $\boldsymbol{s}^\ell \in J$*, we have*

$$f_{\boldsymbol{s}^\ell} - f'_{\boldsymbol{s}^\ell} = Z_{\boldsymbol{s}^\ell} + \Delta_{\boldsymbol{s}^\ell},$$

where $Z_{\boldsymbol{s}^\ell}$ *is a random variable and* $\Delta_{\boldsymbol{s}^\ell}$ *is a deterministic value that satisfy the following:*

- $\hat{\mathbb{E}}\left[Z_{\boldsymbol{s}^\ell}\right] = 0$*;*
- $\hat{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right] \le \mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell} - 2\tilde{m}p(1-p-\alpha)\left|T^\ell_{\boldsymbol{s}^\ell}(i)\cap T'^\ell_{\boldsymbol{s}^\ell}(i)\right|$*;*
- $\hat{\mathbb{P}}\left\{|Z_{\boldsymbol{s}^\ell}| \ge t\right\} \le 2\exp\left(\frac{-t^2/2}{\hat{E}\left[Z^2_{\boldsymbol{s}^\ell}\right]+t/3}\right)$*;*
- $|\Delta_{\boldsymbol{s}^\ell}| \le |\eta_{\boldsymbol{s}^\ell}| + \left|\eta'_{\boldsymbol{s}^\ell}\right| + 2m^{0.2}p$*.*

Furthermore, the random variables $Z_{\boldsymbol{s}^\ell}$ *are conditionally independent for distinct* $\boldsymbol{s}^\ell \in J$*.*

Proof. Lemma G.2 can be proved in a similar way as that of Lemma 5.9 in (Mao et al., 2021a). The main observation behind the proof is that $f_{\boldsymbol{s}^\ell} - f'_{\boldsymbol{s}^\ell} = Z_{\boldsymbol{s}^\ell} + \Delta_{\boldsymbol{s}^\ell}$, where

$$\begin{aligned} Z_{\boldsymbol{s}^\ell} := & \sum_{j \in T^\ell_{\boldsymbol{s}^\ell} \cap T'^\ell_{\boldsymbol{s}^\ell}} \left(\deg_{G(C_k)}(j) - \deg_{G'(C_k)}(j) \right) \\ & + \sum_{j \in T^\ell_{\boldsymbol{s}^\ell} \setminus T'^\ell_J} \left(\deg_{G(C_k)}(j) - 1 - \tilde{m}p \right) \\ & - \sum_{j \in T'^\ell_{\boldsymbol{s}^\ell} \setminus T^\ell_J} \left(\deg_{G'(C_k)}(j) - 1 - \tilde{m}p \right) \end{aligned}$$

and

$$\Delta_{\boldsymbol{s}^\ell} := \eta_{\boldsymbol{s}^\ell} - \eta'_{\boldsymbol{s}^\ell} + (\tilde{m} - m)p \left(\left| T^\ell_{\boldsymbol{s}^\ell} \right| - \left| T'^\ell_{\boldsymbol{s}^\ell} \right| \right).$$

□

Lemma G.3 (Upper bound on the normalized distance of sparsified signature vectors for a correct pair)**.** *For any constants $C, D, K, \delta > 0$, there exist constant ϵ and m_0 with the properties below. Suppose that $m \geq m_0,\ \alpha \in (0, \epsilon),\ w \geq (\log m)^4$, and*

$$(\log m)^{1+\delta} \leq m \frac{p}{1-\alpha} \leq m^{1/20},$$
$$2 \log \left(w^4 \log m \right) \leq k'\ell \leq C \log \log m.$$

Moreover, suppose that a vertex $i \in C_k$ satisfies conditions $(B1) - (B5)$ with a constant $\epsilon > 0$. Consider a subset $J \subset \{-1, 1\}^{k'\ell}$ satisfying $|J| = 2w$. Under the same conditions as stated in Lemma G.2, where $\eta_{\boldsymbol{s}^\ell}$ and $\eta'_{\boldsymbol{s}^\ell}$ satisfy (G.7) and (G.8) with a constant $K > 0$, we have

$$\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i) \right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i)} \leq 2w \left(1 - \frac{1}{(\log m)^{0.1}} \right),$$

with a conditional probability of at least $1 - m^{-D}$.

Lemma G.3 will be proved in Section G.2.

Proof of Lemma E.1. First, we will show that conditions $(B1) - (B5)$ hold for at least $m - m^{1-c}$ vertices $i \in C_k$ with probability at least $1 - m^{-D-1}$.

- By Proposition F.6, there exist at least $m - m^{1-c_1}$ vertices $i \in C_k$ for a constant $c_1 \in (0, 0.5)$ that satisfy condition $(B1)$ with probability at least $1 - m^{-D-2}$.
- Based on Lemma F.2, we have $|\mathcal{B}_{G_0(C_k)}(i, \ell)| = O\left((mp/(1-\alpha))^\ell \right) \leq m^{0.1}$ for all $i \in C_k$ and $\ell \leq \frac{\log m}{11 \log mp}$, with probability at least $1 - m^{-D-2}$. Thus, the condition $(B2)$ holds for all $i \in C_k$ with probability at least $1 - m^{-D-2}$.
- On the conditions $(B1)$ and $(B2)$, the condition $(B3)$ holds for all $i \in C_k$ with probability at least $1 - m^{-D-2}$ by Lemma F.3.
- Condition $(B4)$ holds for all $i \in C_k$ satisfying conditions $(B1)$ and $(B3)$ with probability at least $1 - m^{-D-2}$ by Lemma G.4.
- By Proposition F.6, $|\operatorname{Typ}_{G_0(C_k)}(\ell, \epsilon)| \geq m - m^{1-c_2}$ holds for a constant $c_2 \in (0, 0.5)$ with probability $1 - m^{-D-2}$. Thus, condition $(B5)$ holds for at least $m - m^{1-c_2}$ vertices $i \in C_k$ with probability at least $1 - 2m^{-D-2}$ by Lemma G.6.

Moreover, by applying Lemma G.1 to the vertex i satisfying conditions $(B1) - (B5)$, we obtain $\eta_{\boldsymbol{s}^\ell}$ and $\eta'_{\boldsymbol{s}^\ell}$ that satisfy (G.7) and (G.8) with a probability of at least $1 - m^{-D-2}$. Therefore, by applying Lemma G.3, the proof is complete.

□

G.2. Proof of Lemma G.3

By condition $(B5)$, we have that

$$|T^{\ell}_{\boldsymbol{s}^\ell}(i) \cap T'^{\ell}_{\boldsymbol{s}^\ell}(i)| \geq (mp/2^{k'})^{\ell}(1-6\epsilon)^{k'\ell} \tag{G.11}$$

for any $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$. Thus, we obtain

$$\begin{aligned} \mathrm{v}_{\boldsymbol{s}^\ell} &= mp(1-p)\left|T^{\ell}_{\boldsymbol{s}^\ell}(i)\right| \geq mp(1-p)|T^{\ell}_{\boldsymbol{s}^\ell}(i) \cap T'^{\ell}_{\boldsymbol{s}^\ell}(i)| \\ &\geq (1-p)\left(\frac{1-6\epsilon}{2}\right)^{k'\ell}(mp)^{\ell+1}. \end{aligned} \tag{G.12}$$

Moreover, by condition $(B3)$,

$$\mathrm{v}_{\boldsymbol{s}^\ell} = mp(1-p)\left|T^{\ell}_{\boldsymbol{s}^\ell}(i)\right| \leq 6(1-p)\frac{(mp)^{\ell+1}}{2^{k'\ell}}. \tag{G.13}$$

In the same way, the bounds for $\mathrm{v}'_{\boldsymbol{s}^\ell}$ can be obtained as follows:

$$(1-p)\left(\frac{1-6\epsilon}{2}\right)^{k'\ell}(mp)^{\ell+1} \leq \mathrm{v}'_{\boldsymbol{s}^\ell} \leq 6(1-p)\frac{(mp)^{\ell+1}}{2^{k'\ell}}. \tag{G.14}$$

One can see from Lemma G.2 that for a correct pair of vertices the difference between the entries of signatures $f_{\boldsymbol{s}^\ell}$ and $f'_{\boldsymbol{s}^\ell}$ at some $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$ can be decomposed into the random variable part $Z_{\boldsymbol{s}^\ell}$ and the deterministic part $\Delta_{\boldsymbol{s}^\ell}$. Define

$$X_{\boldsymbol{s}^\ell} := \frac{f_{\boldsymbol{s}^\ell} - f'_{\boldsymbol{s}^\ell}}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}}} = \frac{Z_{\boldsymbol{s}^\ell} + \Delta_{\boldsymbol{s}^\ell}}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}}},$$

where $Z_{\boldsymbol{s}^\ell}$ and $\Delta_{\boldsymbol{s}^\ell}$ are defined in Lemma G.2.

We first estimate the expectation and variance of $X_{\boldsymbol{s}^\ell}$. Recall $\hat{\mathbb{E}}$ defined in Lemma G.2. By Lemma G.2, (G.7),(G.8),(G.12) and (G.13), we can have

$$\begin{aligned} \left|\hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right| = \frac{|\Delta_{\boldsymbol{s}^\ell}|}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}}} &\leq \frac{|\eta_{\boldsymbol{s}^\ell}| + \left|\eta'_{\boldsymbol{s}^\ell}\right| + 2m^{0.2}p}{\sqrt{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}}} \\ &\leq \frac{2K\frac{(mp)^{(\ell+1)/2}}{2^{k'\ell}}w^2\sqrt{\log m} + 2m^{0.2}p}{\sqrt{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}}} \leq \frac{5Kw^2\sqrt{\log m}}{2^{k'\ell/2}(1-6\epsilon)^{k'\ell/2}} \quad \text{for } \boldsymbol{s}^\ell \in \tilde{J}(i), \end{aligned} \tag{G.15}$$

$$\left|\hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right| = \frac{|\Delta_{\boldsymbol{s}^\ell}|}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}}} \leq \frac{2K\frac{(mp)^{(\ell+1)/2}}{2^{k'\ell/2}}\sqrt{\log m} + 2m^{0.2}p}{\sqrt{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}}} \leq \frac{5K\sqrt{\log m}}{(1-6\epsilon)^{k'\ell/2}} \quad \text{for } \boldsymbol{s}^\ell \in J, \tag{G.16}$$

and

$$\begin{aligned} \mathrm{Var}\left(X_{\boldsymbol{s}^\ell}\right) = \frac{\hat{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right]}{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}} &\leq \frac{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell} - 2\tilde{m}p(1-p-\alpha)\left|T^{\ell}_{\boldsymbol{s}^\ell} \cap T'^{\ell}_{\boldsymbol{s}^\ell}\right|}{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}} \\ &\leq 1 - \frac{2\tilde{m}p(1-p-\alpha)(mp/2^{k'})^{\ell}(1-6\epsilon)^{k'\ell}}{12(1-p)(mp)^{\ell+1}/2^{k'\ell}} \leq 1 - \frac{(1-6\epsilon)^{k'\ell}}{13}, \end{aligned} \tag{G.17}$$

if $\alpha \leq \frac{1}{10}$. By (G.15) and (G.17), we can get

$$\hat{\mathbb{E}}\left[X^2_{\boldsymbol{s}^\ell}\right] = \frac{\hat{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right] + \Delta^2_{\boldsymbol{s}^\ell}}{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}} \leq 1 - \frac{(1-6\epsilon)^{k'\ell}}{13} + \frac{25K^2w^4\log m}{2^{k'\ell}(1-6\epsilon)^{k'\ell}} \leq 1 - \frac{(1-6\epsilon)^{k'\ell}}{14} \quad \text{for } \boldsymbol{s}^\ell \in \tilde{J}(i) \tag{G.18}$$

if $2\log\left(w^4\log m\right) \leq k'\ell$, ϵ is sufficiently small and $m \geq m_0(K,\epsilon)$. Similarly, we have

$$\hat{\mathbb{E}}\left[X^2_{\boldsymbol{s}^\ell}\right] = \frac{\hat{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right] + \Delta^2_{\boldsymbol{s}^\ell}}{\mathrm{v}_{\boldsymbol{s}^\ell} + \mathrm{v}'_{\boldsymbol{s}^\ell}} \leq 1 - \frac{(1-6\epsilon)^{k'\ell}}{14} + \frac{25K^2\log m}{(1-6\epsilon)^{k'\ell}} \leq \frac{26K^2\log m}{(1-6\epsilon)^{k'\ell}} \quad \text{for } \boldsymbol{s}^\ell \in J, \tag{G.19}$$

if $k'\ell \le C\log\log m$, ϵ is a sufficiently small constant that depends on C, and $m \ge m_0(K,\epsilon)$. Moreover, we can obtain $\hat{\mathbb{P}}\left\{\left|X_{\boldsymbol{s}^\ell} - \hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right| \ge t\right\} \le 2\exp\left(\frac{-t^2/2}{1+t/3}\right)$ since $\hat{\mathbb{P}}\left\{|Z_{\boldsymbol{s}^\ell}| \ge t\right\} \le 2\exp\left(\frac{-t^2/2}{\hat{E}\left[Z_{\boldsymbol{s}^\ell}^2\right]+t/3}\right)$ holds by Lemma G.2 and $\mathrm{Var}\left(X_{\boldsymbol{s}^\ell}\right) \le 1$ from (G.17). Therefore, applying Lemma I.9, we obtain

$$\sum_{\boldsymbol{s}^\ell \in J} X_{\boldsymbol{s}^\ell}^2 \le \sum_{\boldsymbol{s}^\ell \in J} \hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}^2\right] + C_1\sqrt{w}(\log m)^{3/2} + C_1\left(\max_{\boldsymbol{s}^\ell \in J}\left|\hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right|\right)(\sqrt{w\log m} + \log m), \tag{G.20}$$

with conditional probability at least $1 - m^{-D}$, for $J \subset \{-1,1\}^{k'\ell}$ such that $|J| = 2w$ and a sufficiently large constant C_1 depending on D. By (G.19), (G.18) and (G.1) in condition $(B4)$, we can get

$$\begin{aligned}\sum_{\boldsymbol{s}^\ell \in J} \hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}^2\right] &= \sum_{\boldsymbol{s}^\ell \in \tilde{J}(i)} \hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}^2\right] + \sum_{\boldsymbol{s}^\ell \in J\setminus\tilde{J}(i)} \hat{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}^2\right] \\ &\le 2w\left(1 - \frac{(1-6\epsilon)^{k'\ell}}{14}\right) + (\log m)^2\frac{26K^2\log m}{(1-6\epsilon)^{k'\ell}} \le 2w\left(1 - \frac{(1-6\epsilon)^{k'\ell}}{15}\right)\end{aligned} \tag{G.21}$$

if $k'\ell \le C\log\log m$, $\epsilon > 0$ is a sufficiently small constant that depends on C, $w \ge (\log m)^4$ and $m \ge m_0(K,\epsilon)$. Finally, by (G.16), (G.20) and (G.21), we obtain

$$\begin{aligned}\sum_{\boldsymbol{s}^\ell \in J} X_{\boldsymbol{s}^\ell}^2 &\le 2w\left(1 - \frac{(1-6\epsilon)^{k'\ell}}{15}\right) + C_1\sqrt{w}(\log m)^{3/2} + C_1\frac{5K\sqrt{\log m}}{(1-6\epsilon)^{k'\ell/2}}(\sqrt{w\log m} + \log m) \\ &\le 2w\left(1 - \frac{(1-6\epsilon)^{k'\ell}}{16}\right) \le 2w\left(1 - \frac{1}{(\log m)^{0.1}}\right)\end{aligned}$$

if $k'\ell \le C\log\log m$, ϵ is a sufficiently small constant that depends on C, $w \ge (\log m)^4$ and $m \ge m_0(K,\epsilon)$.

G.3. Proof of Condition $(B4)$

In this subsection, we will show that event $\mathcal{G}$ (or condition $(B4)$) holds with high probability. The idea of comparing sparsified signature vectors was addressed and used in (Mao et al., 2021a;b) to mitigate the dependency issue between the signature vectors. We will follow the same trick here.

Lemma G.4 (Similar to Lemma 5.6 in (Mao et al., 2021a))**.** *Consider a fixed $i \in C_k$ and a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$, for a positive integer w satisfying $w > 2(\log m)^2$. If a vertex i satisfies that*

$(C1)$ $G_0\left(\mathcal{B}_{G_0(C_k)}(i,\ell)\right)$ *forms a tree;*

$(C2)$ $\left|T_{\boldsymbol{s}^\ell}^\ell(i)\right| \vee \left|T_{\boldsymbol{s}^\ell}'^\ell(i)\right| \le 6\left(\frac{mp}{2^{k'}}\right)^\ell$ *for all* $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$,

then event $\mathcal{G}$ (or condition $(B4)$) holds with probability at least $1 - \exp\left(-(\log m)^{1.5}\right)$.

Proof. By the condition $(C2)$, we have

$$\left|T_{\boldsymbol{s}^\ell}^\ell(i)\right| \vee \left|T_{\boldsymbol{s}^\ell}'^\ell(i)\right| \le 6\frac{(mp)^\ell}{2^{k'\ell}} \tag{G.22}$$

for all $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$. Thus, it is easy to see that (G.3) is established since $R_{\boldsymbol{s}^\ell}(i) \subset T_{\boldsymbol{s}^\ell}^\ell(i)$ and $R'_{\boldsymbol{s}^\ell}(i) \subset T_{\boldsymbol{s}^\ell}'^\ell(i)$.

Next, we apply Lemma I.8 with $\Omega = \mathcal{S}_{G(C_k)}(i,\ell), \Omega' = \mathcal{S}_{G'(C_k)}(i,\ell), k = 2^{k'\ell}$, $S = 6(mp)^\ell, L = 8ew^3$, and $\rho = \frac{1}{4w}(\log m)^2$. Moreover, we use $T_{\boldsymbol{s}^\ell}^\ell(i)$ and $T_{\boldsymbol{s}^\ell}'^\ell(i)$ for the partition set of $\Omega = \mathcal{S}_{G(C_k)}(i,\ell)$ and $\Omega' = \mathcal{S}_{G'(C_k)}(i,\ell)$. Then, we can get

$$\left|\left\{\boldsymbol{s}^\ell \in J : \exists \boldsymbol{t}^\ell \in J\backslash\{\boldsymbol{s}^\ell\} \text{ s.t. } \left|T_{\boldsymbol{s}^\ell}^\ell(i) \cap T_{\boldsymbol{t}^\ell}'^\ell(i)\right| \ge 48ew^3\frac{(mp)^\ell}{4^{k'\ell}}\right\}\right| < \frac{1}{2}(\log m)^2, \tag{G.23}$$

with probability at least $1 - \exp\left(-(\log m)^2/4\right)$. In a similar way, the following results can be obtained.

$$\left|\left\{\boldsymbol{s}^\ell \in J : \exists \boldsymbol{t}^\ell \in J\backslash\{\boldsymbol{s}^\ell\} \text{ s.t. } \left|T_{\boldsymbol{s}^\ell}'^\ell(i) \cap T_{\boldsymbol{t}^\ell}^\ell(i)\right| \ge 48ew^3\frac{(mp)^\ell}{4^{k'\ell}}\right\}\right| < \frac{1}{2}(\log m)^2, \tag{G.24}$$

with probability at least $1-\exp\left(-(\log m)^2/4\right)$. Define

$$\tilde{J}(i) := \left\{ \boldsymbol{s}^\ell \in J : |R_{\boldsymbol{s}^\ell}(i)| \vee |R'_{\boldsymbol{s}^\ell}(i)| \le 96ew^4 \frac{(mp)^\ell}{4^{k'\ell}} \right\}.$$

We can see that $\tilde{J}(i)$ is a superset of

$$\left\{ \boldsymbol{s}^\ell \in J : \forall \boldsymbol{t}^\ell \in J \backslash \{\boldsymbol{s}^\ell\},\ \left|T^\ell_{\boldsymbol{s}^\ell}(i) \cap T'^\ell_{\boldsymbol{t}^\ell}(i)\right| \le 48ew^3 \frac{(mp)^\ell}{4^{k'\ell}} \right\}$$
$$\bigcap \left\{ \boldsymbol{s}^\ell \in J : \forall \boldsymbol{t}^\ell \in J \backslash \{\boldsymbol{s}^\ell\},\ \left|T'^\ell_{\boldsymbol{s}^\ell}(i) \cap T^\ell_{\boldsymbol{t}^\ell}(i)\right| \le 48ew^3 \frac{(mp)^\ell}{4^{k'\ell}} \right\}.$$

By combining the above result with (G.23) and (G.24), we see that (G.1) and (G.2) hold. By taking a union bound over all $i \in C_k$, we can complete the proof. □

G.4. Proof of Condition $(B5)$

In this subsection, we will show that 'typical' vertices, defined Definition F.4, satisfy condition $(B5)$ with high probability, i.e., $\left|T^\ell_{\boldsymbol{s}^\ell}(i) \cap T'^\ell_{\boldsymbol{s}^\ell}(i)\right| \ge (mp/2^{k'})^\ell (1-6\epsilon)^{k'\ell}$ for a constant ϵ and for all $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$ with high probability.

Lemma G.5 (Degree correlation between correct pairs of vertices)**.** *For any constant $\epsilon > 0$, there exist constants $\alpha_0, L > 0$ depending only on ϵ with the properties below. Consider the two graphs G^π and G', which are generated from the correlated SBMs defined in Sec. 1.1, with correlation $1-\alpha$. Suppose that community labels $(C_1, C_2, \ldots, C_k)$ are given in both graphs and $n_a q \ge L$. If $\alpha \in (0, \alpha_0)$, then for all $i \in C_k$ and any $a \in [k-1]$,*

$$\mathbb{P}\left\{\mathbf{Sign}(\deg^a_G(i) - n_a q) = \mathbf{Sign}(\deg^a_{G'}(i) - n_a q)\right\} \ge 1-\epsilon. \tag{G.25}$$

Proof. From Lemma I.2, we can show that there exists a universal constant $C > 0$ such that for any $r > 0$,

$$\mathbb{P}\left\{ |\deg^a_G(i) - n_a q| > r\sqrt{n_a q(1-q)} \right\} \ge 1 - Cr. \tag{G.26}$$

By applying Lemma I.7 with $J = J' = C_a$, for any $t > 0$, we have

$$\mathbb{P}\left\{ |\deg^a_G(i) - \deg^a_{G'}(i)| \ge 4\left(t + \sqrt{t\alpha q n_a}\right) \right\} \le 6\exp(-t) + 2\exp\left(-\frac{q n_a}{3(1-\alpha)}\right). \tag{G.27}$$

Let $t = \frac{r\sqrt{n_a q(1-q)}}{8} \wedge \frac{r^2(1-q)}{64\alpha}$. Then, we can get

$$4\left(t + \sqrt{t\alpha q n_a}\right) \le r\sqrt{n_a q(1-q)}. \tag{G.28}$$

From (G.26) and (G.27), we have

$$\begin{aligned}
&\mathbb{P}\left\{\mathbf{Sign}(\deg^a_G(i) - n_a q) = \mathbf{Sign}(\deg^a_{G'}(i) - n_a q)\right\} \\
&\ge \mathbb{P}\left\{ |\deg^a_G(i) - n_a q| > r\sqrt{n_a q(1-q)} \text{ and } |\deg^a_G(i) - \deg^a_{G'}(i)| < r\sqrt{n_a q(1-q)} \right\} \\
&\ge 1 - Cr - 6\exp(-t) - 2\exp\left(-\frac{q n_a}{3(1-\alpha)}\right).
\end{aligned} \tag{G.29}$$

We can take a small enough r to make $Cr < \epsilon/3$. We can also make $2\exp\left(-\frac{q n_a}{3(1-\alpha)}\right) < \epsilon/3$ and $6\exp(-t) < \epsilon/3$ by setting $t = \frac{r\sqrt{n_a q(1-q)}}{8} \wedge \frac{r^2(1-q)}{64\alpha}$, since $n_a q \ge L$ for a sufficiently large constant L depending on ϵ and we can take a small enough α_0. Thus, the proof is complete. □

We will prove condition $(B5)$ using the above results. Similar to Proposition 5.1 in (Mao et al., 2021a), we prove that there exists a significant overlap between the partitioning nodes of a correct pair.

Proposition G.6 (Overlap between the partitioning nodes of a correct pair)**.** *For any constants $D, \delta, \epsilon > 0$, there exist constants m_0, Q, k_1 and α_1, which depend only on D, δ and ϵ, with the properties below. Consider the two graphs G and G', which are generated from the correlated SBMs defined in Sec. 1.1, with correlation $1-\alpha$. Suppose that community labels $(C_1, C_2, \dots, C_k)$ are given. Assume a given instance of G_0. For a fixed positive integer ℓ, assume that*

$$m \geq m_0, \quad (\log m)^{1+\delta} \leq mp/(1-\alpha) \leq m^{1/20}, \quad k' \leq k_1 \log mp, \quad mq \geq Qk'^2, \quad \alpha \in (0, \alpha_1).$$

Then, for every vertex $i \in \mathrm{Typ}_{G_0(C_k)}(\ell, \epsilon)$, satisfying Definition F.4, and for any $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$,

$$\left|T^\ell_{\boldsymbol{s}^\ell}(i, G) \cap T^\ell_{\boldsymbol{s}^\ell}(i, G')\right| \geq \left(\frac{mp}{2^{k'}}\right)^\ell (1-6\epsilon)^{k'\ell},$$

with probability at least $1 - m^{-D}$.

Proof. Let $\mathbb{P}_0$ represent the conditional probability given an instance of G_0. Fix a vertex $i \in \mathrm{Typ}_{G_0(C_k)}(\ell, \epsilon)$, and let $d \in \{0, 1, \dots, \ell-1\}$. For any $\epsilon > 0$ and $j \in \mathcal{S}_{G_0(C_k)}(i, d)$, we have

$$\begin{aligned}
&\mathbb{P}_0\{\deg_{G\cap G'(C_k)}(j) \leq (1-3\epsilon)mp\} \\
&\leq \mathbb{P}_0\{(1-\alpha)^2 \deg_{G_0(C_k)}(j) - \deg_{G\cap G'(C_k)}(j) > (1-\alpha)^2 \deg_{G_0(C_k)}(j) - (1-3\epsilon)mp\} \\
&\overset{(a)}{\leq} \mathbb{P}_0\{(1-\alpha)^2 \deg_{G_0(C_k)}(j) - \deg_{G\cap G'(C_k)}(j) > \epsilon mp\} \overset{(b)}{\leq} \exp\left(-Kmp\right) \leq m^{-D-2},
\end{aligned} \tag{G.30}$$

where (a) holds from condition $(A3)$ and by choosing $\alpha < \epsilon$. We have that $\deg_{G_0(C_k)}(j) \leq 2m\frac{p}{1-\alpha}$ by condition $(A2)$ and $\deg_{G\cap G'(C_k)}(j)$ is Binomial $\left(\deg_{G_0(C_k)}(j), (1-\alpha)^2\right)$. Therefore, using Bernstein's inequality (Lemma I.5), it can be confirmed that the inequality (b) holds for a constant K depending only on ϵ. The last inequality holds since $mp/(1-\alpha) \geq (\log m)^{1+\delta}$, $\alpha < \epsilon$ and $m \geq m_0(D, \delta, \epsilon)$.

Define an event

$$\mathcal{F}(i, d) := \left\{\deg_{G\cap G'(C_k)}(j) \geq (1-3\epsilon)mp, \ \ \forall j \in \mathcal{S}_{G_0(C_k)}(i, d)\right\}. \tag{G.31}$$

By applying a union bound over $j \in \mathcal{S}_{G_0(C_k)}(i, d)$ and $d \in \{0, \dots, \ell-1\}$, we can obtain

$$\mathbb{P}_0\left\{\cap_{d=0}^{\ell-1}\mathcal{F}(i, d)\right\} \geq 1 - m^{-D-1}. \tag{G.32}$$

On the event $\cap_{d=0}^{\ell-1}\mathcal{F}(i, d)$, we will prove that

$$\left|T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right| \geq \left(\frac{mp}{2^{k'}}\right)^d (1-6\epsilon)^{k'd} \tag{G.33}$$

for all $d \in \{0, 1, \dots, \ell\}$ and $\boldsymbol{s}^d \in \{-1,1\}^{k'd}$. The case where $d = 0$ holds trivially. Assume that (G.33) holds for $d \in \{0, 1, \dots, \ell-1\}$ and $\boldsymbol{s}^d \in \{-1,1\}^{k'd}$. For any $\boldsymbol{s}^d \in \{-1,1\}^{k'd}$ and $\boldsymbol{s}_{d+1} \in \{-1,1\}^{k'}$, let $\boldsymbol{s}^{d+1} = (\boldsymbol{s}^d, \boldsymbol{s}_{d+1}) \in \{-1,1\}^{k'(d+1)}$. On the event $\cap_{d=0}^{\ell-1}\mathcal{F}(i, d)$, we also have that

$$\left|\mathcal{N}_{G\cap G'(C_k)}\left(T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right) \cap \mathcal{S}_{G_0(C_k)}(i, d+1)\right| \geq (1-4\epsilon)mp\left|\left(T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right)\right| \tag{G.34}$$

because of the condition $(A1)$. For any $j \in \mathcal{N}_{G\cap G'(C_k)}\left(T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right) \cap \mathcal{S}_{G_0(C_k)}(i, d+1)$, we have

$$\mathbb{P}_0\left\{j \in T^{d+1}_{\boldsymbol{s}^{d+1}}(i, G) \cap T^{d+1}_{\boldsymbol{s}^{d+1}}(i, G')\right\} \geq \frac{(1-\epsilon)^{k'}}{2^{k'}}\left(1 - \frac{2Ck'}{\sqrt{mq}}\right) \geq \frac{(1-\epsilon)^{k'}}{2^{k'}} \cdot (1-\epsilon) \tag{G.35}$$

for a universal constant C and $\alpha \in (0, \alpha_0)$ because of Lemma G.5 and (F.8) with $(1-x)^n \geq 1 - nx$. The last inequality holds from $mq \geq Qk'^2$ for a sufficiently large constant Q depending on ϵ. Therefore, by Hoeffding's inequality (Lemma I.4), we obtain

$$\begin{aligned}
&\mathbb{P}_0\left\{\left|T^{d+1}_{\boldsymbol{s}^{d+1}}(i, G) \cap T^{d+1}_{\boldsymbol{s}^{d+1}}(i, G')\right| \geq \left(\frac{1-6\epsilon}{2}\right)^{k'} mp\left|T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right|\right\} \\
&\geq 1 - \exp\left(-K'\left(\frac{1-6\epsilon}{2}\right)^{2k'} mp\left|T^d_{\boldsymbol{s}^d}(i, G) \cap T^d_{\boldsymbol{s}^d}(i, G')\right|\right) \\
&\geq 1 - m^{-D-1}
\end{aligned} \tag{G.36}$$

for a positive constant K', which depends on ϵ. The last inequality holds since $k' \le k_1 \log mp$ for a constant k_1 depending on ϵ, δ such that $\left(\frac{1-6\epsilon}{2}\right)^{2k_1 \log mp} \ge (\log m)^{-0.5\delta/(1+\delta)}$ and $m \ge m_0(D, \delta, \epsilon)$. Let $\alpha_1 = \alpha_0 \wedge \epsilon$. Taking a union bound over $d \in \{0, 1, \ldots, \ell - 1\}$ and $\boldsymbol{s}^d \in \{-1, 1\}^{k'd}$, we have

$$\left|T^{\ell}_{\boldsymbol{s}^\ell}(i, G) \cap T^{\ell}_{\boldsymbol{s}^\ell}(i, G')\right| \ge \left(\frac{mp}{2^{k'}}\right)^{\ell} (1 - 6\epsilon)^{k'\ell} \tag{G.37}$$

with probability at least $1 - m^{-D}$. □

H. Proof of Lemma E.2

Consider a random subset J uniformly drawn from $\{-1, 1\}^{k'\ell}$ with cardinality $2w$, where w is an integer satisfying $2w > \log m$. We define the event $\mathcal{H}$ to be true if and only if there exist positive constants K and K' such that the following conditions hold:

$$\left(\max_{\boldsymbol{s}^\ell \in J} |L_{\boldsymbol{s}^\ell}(i, i')|\right) \vee \left(\max_{\boldsymbol{s}^\ell \in J} |L'_{\boldsymbol{s}^\ell}(i, i')|\right) \le K \frac{(mp)^\ell}{2^{k'\ell}}, \tag{H.1}$$

$$\left(\sum_{\boldsymbol{s}^\ell \in J} |L_{\boldsymbol{s}^\ell}(i, i')|\right) \vee \left(\sum_{\boldsymbol{s}^\ell \in J} |L'_{\boldsymbol{s}^\ell}(i, i')|\right) \le K \frac{(mp)^\ell}{2^{k'\ell}} \left(\frac{w}{mp(1-\alpha)^{\ell-1}} + \sqrt{w \log m}\right), \tag{H.2}$$

where

$$L_{\boldsymbol{s}^\ell}(i, i') := T^{\ell}_{\boldsymbol{s}^\ell}(i) \cap \mathcal{B}_{G_0(C_k)}(i', \ell), \tag{H.3}$$

$$L'_{\boldsymbol{s}^\ell}(i, i') := T'^{\ell}_{\boldsymbol{s}^\ell}(i') \cap \mathcal{B}_{G_0(C_k)}(i, \ell) \tag{H.4}$$

and

$$\left(\max_{\boldsymbol{s}^\ell \in J} |\zeta_{\boldsymbol{s}^\ell}(i, i')|\right) \vee \left(\max_{\boldsymbol{s}^\ell \in J} |\zeta'_{\boldsymbol{s}^\ell}(i, i')|\right) \le K' \frac{(mp)^{(\ell+1)/2}}{2^{k'\ell/2}} \sqrt{\log m}, \tag{H.5}$$

where

$$\zeta_{\boldsymbol{s}^\ell}(i, i') := \sum_{j \in L_{\boldsymbol{s}^\ell}(i, i')} \left(\deg_{G(C_k)}(j) - 1 - mp\right) \quad \text{and} \quad \zeta'_{\boldsymbol{s}^\ell}(i, i') := \sum_{j \in L'_{\boldsymbol{s}^\ell}(i, i')} \left(\deg_{G'(C_k)}(j) - 1 - mp\right). \tag{H.6}$$

Let us define the four conditions $(D1) - (D4)$ as follows.

$(D1)$ $G_0\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0(C_k)}(i', \ell)\right)$ forms a tree or a forest of two trees.
$(D2)$ $\left|\mathcal{B}_{G_0(C_k)}(i, \ell)\right| + \left|\mathcal{B}_{G_0(C_k)}(i', \ell)\right| \le m^{0.1}$.
$(D3)$ Event $\mathcal{H}$ holds.
$(D4)$ $\left|T^{\ell}_{\boldsymbol{s}^\ell}(i)\right| \wedge \left|T'^{\ell}_{\boldsymbol{s}^\ell}(i')\right| \ge (mp/2^{k'})^\ell (1 - 6\epsilon)^{k'\ell}$, for a sufficiently small positive constant ϵ and for all $\boldsymbol{s}^\ell \in \{-1, 1\}^{k'\ell}$.

We will show that there exists a sufficiently large vertex set $\mathcal{I} \subset C_k$ such that conditions $(D1) - (D4)$ hold for every pair of $i, i' \in \mathcal{I}$, $i \ne i'$, and for all such wrong pairs (E.2) holds with high probability. The conditions $(D1)$, $(D2)$ and $(D4)$ will be shown to be held with high probability by using the analysis from Sec. F and Sec.G.4, and the condition $(D3)$ will be proven in Sec. H.3.

H.1. Proof of Lemma E.2

We consider a fixed distinct pair $i, i' \in C_k$ and a subset J uniformly drawn from $\{-1, 1\}^{k'\ell}$ with $|J| = 2w$. In addition, we condition on the subgraphs

$$G_0\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0(C_k)}(i', \ell)\right), G\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0}(C_k)(i', \ell)\right), G'\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0(C_k)}(i', \ell)\right)$$

as well as every edge between

$$\mathcal{S}_{G_0(C_k)}(i', \ell) \cap \mathcal{B}_{G_0(C_k)}(i, \ell) \text{ and } \mathcal{S}_{G_0(C_k)}(i', \ell + 1)$$

in $G_0(C_k)$, $G(C_k)$ and $G'(C_k)$, and also every edge between

$$\mathcal{S}_{G_0(C_k)}(i,\ell) \cap \mathcal{B}_{G_0(C_k)}(i',\ell) \text{ and } \mathcal{S}_{G_0(C_k)}(i,\ell+1)$$

in graphs $G_0(C_k)$, $G(C_k)$ and $G'(C_k)$ such that conditions $(D1)-(D4)$ hold. Under these conditions, $\zeta_{\boldsymbol{s}^\ell}(i,i')$ and $\zeta'_{\boldsymbol{s}^\ell}(i,i')$ are deterministic quantities defined in (H.6). Furthermore, we assume that the remaining edges are randomly generated. Denote the conditional probability and expectation as $\overline{\mathbb{P}}$ and $\overline{\mathbb{E}}$, respectively. By conditions $(D1)$ and $(D2)$, we can see that $\deg_{G(C_k)}(j)-1$ follows a binomial distribution with parameters $\bar{m}$ and p, where

$$\bar{m} := m - \left|\mathcal{B}_{G_0(C_k)}(i,\ell) \cup \mathcal{B}_{G_0(C_k)}(i',\ell)\right| \geq m - m^{0.1},$$

for any $j \in \mathcal{S}_{G_0(C_k)}(i,\ell)\backslash\mathcal{B}_{G_0(C_k)}(i',\ell)$. The same statement holds for $\deg_{G'(C_k)}(j)-1$. Furthermore, it is independent for each $j \in \mathcal{S}_{G_0(C_k)}(i,\ell)\backslash\mathcal{B}_{G_0(C_k)}(i',\ell)$. Since vertices i and i' are fixed, we omit (i,i') from $\zeta_{\boldsymbol{s}^\ell}(i,i')$ and $\zeta'_{\boldsymbol{s}^\ell}(i,i')$ defined in (H.6) for notational simplicity. We also use $L_{\boldsymbol{s}^\ell}$ and $L'_{\boldsymbol{s}^\ell}$ for the quantities defined in (H.3) and (H.4).

Lemma H.1 (Similar to Lemma 6.3 in (Mao et al., 2021a))**.** *For each $\boldsymbol{s}^\ell \in J$, we can express the difference between $f_{\boldsymbol{s}^\ell}(i)$ and $f'_{\boldsymbol{s}^\ell}(i')$ as*

$$f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i') = Z_{\boldsymbol{s}^\ell} + \Delta_{\boldsymbol{s}^\ell}$$

where $Z_{\boldsymbol{s}^\ell}$ is a random variable and $\Delta_{\boldsymbol{s}^\ell}$ is a deterministic value that satisfy the following:

- $\overline{\mathbb{E}}\left[Z_{\boldsymbol{s}^\ell}\right] = 0$;
- $\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i') - 2m^{0.2}p - p(1-p)\bar{m}\left(\left|L_{\boldsymbol{s}^\ell}\right| + \left|L'_{\boldsymbol{s}^\ell}\right|\right) \leq \overline{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right] \leq \mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')$;
- $\overline{\mathbb{P}}\left\{\left|Z_{\boldsymbol{s}^\ell}\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2/2}{E\left[Z^2_{\boldsymbol{s}^\ell}\right]+t/3}\right)$;
- $\left|\Delta_{\boldsymbol{s}^\ell}\right| \leq \left|\zeta_{\boldsymbol{s}^\ell}\right| + \left|\zeta'_{\boldsymbol{s}^\ell}\right| + 2m^{0.2}p$.

Furthermore, the random variables $Z_{\boldsymbol{s}^\ell}$ are conditionally independent for different $\boldsymbol{s}^\ell \in J$.

Proof. The proof of Lemma H.1 follows a similar approach to that of Lemma 6.3 in (Mao et al., 2021a). The key observation is that we can express $f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')$ as the sum of $Z_{\boldsymbol{s}^\ell}$ and $\Delta_{\boldsymbol{s}^\ell}$, where

$$\begin{aligned} Z_{\boldsymbol{s}^\ell} := & \sum_{j \in T^\ell_{\boldsymbol{s}^\ell}(i)\backslash\mathcal{B}_{G_0(C_k)}(i',\ell)} \left(\deg_{G(C_k)}(j) - 1 - \bar{m}p\right) \\ & - \sum_{j \in T'^\ell_{\boldsymbol{s}^\ell}(i')\backslash\mathcal{B}_{G_0(C_k)}(i,\ell)} \left(\deg_{G'(C_k)}(j) - 1 - \bar{m}p\right) \end{aligned}$$

and

$$\Delta_{\boldsymbol{s}^\ell} := \zeta_{\boldsymbol{s}^\ell} - \zeta'_{\boldsymbol{s}^\ell} + (\bar{m}-m)p\left(\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right| - \left|T'^\ell_{\boldsymbol{s}^\ell}(i')\right|\right).$$

□

Lemma H.2 (Lower bound on the normalized distance of sparsified signature vectors for a wrong pair)**.** *For any constants $C, D, K, K', \delta > 0$, there exist constants $m_0, \epsilon > 0$ with the properties below. Let $m \geq m_0$, $\alpha \in (0,\epsilon)$, $w \geq \lfloor(\log m)^5\rfloor$, and*

$$(\log m)^{1+\delta} \leq m\frac{p}{1-\alpha} \leq m^{1/20},\ k'\ell \leq C\log\log m.$$

Furthermore, let conditions $(D1)-(D4)$ hold with constants $K, K', \epsilon > 0$, and consider a subset J of cardinality $2w$. Then, with a conditional probability larger than $1-m^{-D}$, the following statement holds:

$$\sum_{\boldsymbol{s}^\ell \in J} \frac{\left(f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')\right)^2}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')} \geq 2w\left(1 - \frac{1}{(\log m)^{0.9}}\right).$$

Lemma H.2 will be proven in Section H.2.

Proof of Lemma E.2. First, we will establish the existence of a subset $\mathcal{I} \subset C_k$ of cardinality $|\mathcal{I}| > m - m^{1-c}$ such that for any $i, i' \in \mathcal{I}$, $i \neq i'$, conditions $(D1) - (D4)$ are satisfied with a probability of at least $1 - m^{-D-1}$.

- Using Proposition F.6, we can deduce that with a probability of at least $1 - m^{-D-2}$ (by choosing a sufficiently large constant R), there exists a subset $\mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)$ such that $|\mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)| \geq m - m^{1-c_1}$, where $c_1 \in (0, 0.5)$ is a constant. For any distinct vertices i and i' in $\mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)$, condition $(D1)$ holds because if $\mathrm{dist}_{G_0(C_k)}(i, i') \leq 2\ell$, $G_0\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0(C_k)}(i', \ell)\right)$ becomes a forest, and if $\mathrm{dist}_{G_0(C_k)}(i, i') > 2\ell$, $G_0\left(\mathcal{B}_{G_0(C_k)}(i, \ell) \cup \mathcal{B}_{G_0(C_k)}(i', \ell)\right)$ becomes two trees.
- By Lemma F.2, we have $|\mathcal{B}_{G_0(C_k)}(i, \ell)| = O\left((mp)^\ell\right) \leq \frac{1}{2} m^{0.1}$ for all $i \in C_k$ and $\ell \leq \frac{\log m}{11 \log mp}$, with probability at least $1 - m^{-D-2}$. Thus, the condition $(D2)$ holds for all $i \in C_k$ with probability at least $1 - m^{-D-2}$.
- We can assert that the condition $(D3)$ holds with $\mathcal{I} \subset C_k$ of cardinality at least $m - m^{1-c_2}$ for a constant $c_2 \in (0, 0.5)$ with probability at least $1 - m^{-D-2}$ by Lemma H.4.
- By Proposition F.6, $|\mathrm{Typ}_{G_0(C_k)}(\ell, \epsilon)| \geq m - m^{1-c_3}$ for a constant $c_3 \in (0, 0.5)$ with probability at least $1 - m^{-D-2}$. Thus, the condition $(D4)$ holds for at least $m - m^{1-c_3}$ vertices $i \in C_k$ with probability at least $1 - 2m^{-D-2}$ by Lemma G.6 since $\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right| \geq \left|T^\ell_{\boldsymbol{s}^\ell}(i, G) \cap T^\ell_{\boldsymbol{s}^\ell}(i, G')\right|$.

The proof is complete by applying Lemma H.2 . □

H.2. Proof of Lemma H.2

Proof of Lemma H.2. For fixed distinct vertices $i, i' \in C_k$, by condition $(D4)$, we have

$$|T^\ell_{\boldsymbol{s}^\ell}(i)| \wedge |T'^\ell_{\boldsymbol{s}^\ell}(i')| \geq (mp/2^{k'})^\ell (1 - 6\epsilon)^{k'\ell} \tag{H.7}$$

for any $\boldsymbol{s}^\ell \in \{-1, 1\}^{k'\ell}$. Thus, we obtain

$$\begin{aligned} \mathrm{v}_{\boldsymbol{s}^\ell}(i) &= mp(1-p)\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right| \geq mp(1-p)\left(|T^\ell_{\boldsymbol{s}^\ell}(i)| \wedge |T'^\ell_{\boldsymbol{s}^\ell}(i')|\right) \\ &\geq (1-p)\left(\frac{1-6\epsilon}{2}\right)^{k'\ell}(mp)^{\ell+1}. \end{aligned} \tag{H.8}$$

Similarly, we obtain

$$\mathrm{v}'_{\boldsymbol{s}^\ell}(i') \geq (1-p)\left(\frac{1-6\epsilon}{2}\right)^{k'\ell}(mp)^{\ell+1}. \tag{H.9}$$

One can see from Lemma H.1 that for a wrong pair of vertices the difference between the entries of signatures $f_{\boldsymbol{s}^\ell}(i)$ and $f'_{\boldsymbol{s}^\ell}(i')$ at some $\boldsymbol{s}^\ell \in \{-1, 1\}^{k'\ell}$ can be decomposed into the random variable part $Z_{\boldsymbol{s}^\ell}$ and the deterministic part $\Delta_{\boldsymbol{s}^\ell}$. Define

$$X_{\boldsymbol{s}^\ell} := \frac{f_{\boldsymbol{s}^\ell}(i) - f'_{\boldsymbol{s}^\ell}(i')}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')}} = \frac{Z_{\boldsymbol{s}^\ell} + \Delta_{\boldsymbol{s}^\ell}}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')}},$$

where $Z_{\boldsymbol{s}^\ell}$ and $\Delta_{\boldsymbol{s}^\ell}$ are defined in Lemma H.1.

We will first estimate expectation and variance of $X_{\boldsymbol{s}^\ell}$. By Lemma H.1, (H.5), and (H.8), we can have

$$\left|\overline{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right| = \frac{|\Delta_{\boldsymbol{s}^\ell}|}{\sqrt{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')}} \leq \frac{2K' \frac{(mp)^{(\ell+1)/2}}{2^{k'\ell/2}}\sqrt{\log m} + 2m^{0.2}p}{\sqrt{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}}} \leq \frac{6K'\sqrt{\log m}}{(1-6\epsilon)^{k'\ell/2}} \tag{H.10}$$

and

$$\begin{aligned} \mathrm{Var}\left(X_{\boldsymbol{s}^\ell}\right) = \frac{\overline{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right]}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')} &\geq \frac{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i') - 2m^{0.2}p - p(1-p)\bar{m}\left(|L_{\boldsymbol{s}^\ell}| + \left|L'_{\boldsymbol{s}^\ell}\right|\right)}{\mathrm{v}_{\boldsymbol{s}^\ell}(i) + \mathrm{v}'_{\boldsymbol{s}^\ell}(i')} \\ &\geq 1 - \frac{2m^{0.2}p + p(1-p)\bar{m}\left(|L_{\boldsymbol{s}^\ell}| + \left|L'_{\boldsymbol{s}^\ell}\right|\right)}{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}}, \end{aligned} \tag{H.11}$$

for any $\boldsymbol{s}^\ell \in J$. Moreover, we have that $\overline{\mathbb{P}}\left\{\left|X_{\boldsymbol{s}^\ell} - \overline{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2/2}{1+t/3}\right)$ since $\overline{\mathbb{P}}\left\{\left|Z_{\boldsymbol{s}^\ell}\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2/2}{E\left[Z^2_{\boldsymbol{s}^\ell}\right]+t/3}\right)$ and $\mathrm{Var}\left(X_{\boldsymbol{s}^\ell}\right) = \frac{\overline{\mathbb{E}}\left[Z^2_{\boldsymbol{s}^\ell}\right]}{\mathrm{v}_{\boldsymbol{s}^\ell}(i)+\mathrm{v}'_{\boldsymbol{s}^\ell}(i')} \leq 1$ hold from Lemma H.1. Therefore, applying Lemma I.9, we can obtain

$$\sum_{\boldsymbol{s}^\ell \in J} X^2_{\boldsymbol{s}^\ell} \geq \sum_{\boldsymbol{s}^\ell \in J} \overline{\mathbb{E}}\left[X^2_{\boldsymbol{s}^\ell}\right] - C_1\sqrt{w}(\log m)^{3/2} - C_1\left(\max_{\boldsymbol{s}^\ell \in J}\left|\overline{\mathbb{E}}\left[X_{\boldsymbol{s}^\ell}\right]\right|\right)\left(\sqrt{w\log m} + \log m\right) \tag{H.12}$$

with conditional probability at least $1 - m^{-D}$ for $J \subset \{-1,1\}^{k'\ell}$ such that $|J| = 2w$ and a sufficiently large constant C_1 depending on D. By (H.2) in condition $(D3)$, we can get

$$\begin{aligned}
\sum_{\boldsymbol{s}^\ell \in J} \overline{\mathbb{E}}\left[X^2_{\boldsymbol{s}^\ell}\right] &\geq \sum_{\boldsymbol{s}^\ell \in J} \mathrm{Var}(X_{\boldsymbol{s}^\ell}) \geq 2w - \sum_{\boldsymbol{s}^\ell \in J} \frac{2m^{0.2}p + p(1-p)\bar{m}\left(\left|L_{\boldsymbol{s}^\ell}\right| + \left|L'_{\boldsymbol{s}^\ell}\right|\right)}{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}} \\
&\geq 2w - m^{-0.7} - \frac{p(1-p)\bar{m}}{2(1-p)(mp)^{\ell+1}((1-6\epsilon)/2)^{k'\ell}} \cdot 2K\frac{(mp)^\ell}{2^{k'\ell}}\left(\frac{w}{mp(1-\alpha)^{\ell-1}} + \sqrt{w\log m}\right) \\
&\geq 2w\left(1 - \frac{K_1}{(1-7\epsilon)^{k'\ell}\log m}\right)
\end{aligned} \tag{H.13}$$

for a sufficiently large constant K_1 depending on K, if $w \geq (\log m)^3$, $\alpha < \epsilon$, $mp/(1-\alpha) \geq (\log m)^{1+\delta}$ and $m \geq m_0(\epsilon)$. Finally, by (H.10), (H.12) and (H.13), we can obtain

$$\begin{aligned}
\sum_{\boldsymbol{s}^\ell \in J} X^2_{\boldsymbol{s}^\ell} &\geq 2w\left(1 - \frac{K_1}{(1-7\epsilon)^{k'\ell}\log m}\right) - C_1\sqrt{w}(\log m)^{3/2} - C_1\frac{6K'\sqrt{\log m}}{(1-6\epsilon)^{k'\ell/2}}(\sqrt{w\log m} + \log m) \\
&\overset{(a)}{\geq} 2w\left(1 - \frac{K_2}{(1-7\epsilon)^{k'\ell}\log m}\right) \geq 2w\left(1 - \frac{1}{(\log m)^{0.9}}\right).
\end{aligned}$$

The inequality (a) holds by $w \geq \lfloor(\log m)^5\rfloor$ and the last inequality holds since $k'\ell \leq C\log\log m$ and $\epsilon > 0$ is a sufficiently small constant, which depends on C. This completes the proof. □

H.3. Proof of Condition $(D3)$

As in the analysis for correct pairs, we consider the sparsification for a distinct pair of vertices (i, i') in the lemma below (similar to Lemma 6.1 in (Mao et al., 2021a)).

Lemma H.3. *For constants $K_1, D > 0$, there exists $K > 0$ with the properties below. Consider a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$, where w is an integer satisfying $2w \geq \log m$. Then, for any distinct vertices $i, i' \in C_k$ if (i, i') satisfies that*

$(E1)$ $G_0\left(\mathcal{B}_{G_0(C_k)}(i,\ell)\right)$ *and* $G_0\left(\mathcal{B}_{G_0(C_k)}(i',\ell)\right)$ *form trees;*

$(E2)$ $\left|T^\ell_{\boldsymbol{s}^\ell}(i)\right| \vee \left|T'^\ell_{\boldsymbol{s}^\ell}(i')\right| \leq 6\left(\frac{mp}{2^{k'}}\right)^\ell$ *for all* $\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}$;

$(E3)$ $\left|\mathcal{B}_{G_0(C_k)}(i,\ell) \cap \mathcal{B}_{G_0(C_k)}(i',\ell)\right| \leq K_1\left(\frac{mp}{1-\alpha}\right)^{\ell-1}$,

then $L_{\boldsymbol{s}^\ell}(i,i')$ and $L'_{\boldsymbol{s}^\ell}(i,i')$ defined in (H.3) and (H.4) satisfy (H.1) and (H.2) with probability at least $1 - m^{-D}$.

Proof. For fixed vertices $i, i' \in C_k$, $i \neq i'$, let

$$a_{\boldsymbol{s}^\ell} := \left|L_{\boldsymbol{s}^\ell}(i,i')\right|,\ a'_{\boldsymbol{s}^\ell} := \left|L'_{\boldsymbol{s}^\ell}(i,i')\right|.$$

Since $L_{\boldsymbol{s}^\ell}(i,i') \subset T^\ell_{\boldsymbol{s}^\ell}(i)$, $L'_{\boldsymbol{s}^\ell}(i,i') \subset T'^\ell_{\boldsymbol{s}^\ell}(i)$ and from the condition $(E2)$, we get

$$a_{\boldsymbol{s}^\ell} \vee a'_{\boldsymbol{s}^\ell} \leq 6\left(\frac{mp}{2^{k'}}\right)^\ell. \tag{H.14}$$

Therefore, (H.1) holds. By conditions $(E1)$ and $(E3)$, we can obtain that

$$\sum_{\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}} a_{\boldsymbol{s}^\ell} \le \left|\mathcal{B}_{G_0(C_k)}(i,\ell) \cap \mathcal{B}_{G_0(C_k)}(i',\ell)\right| \le K_1 \left(\frac{mp}{1-\alpha}\right)^{\ell-1}.$$

For a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w \ge 2\log m$, by applying Lemma I.6 we have

$$\begin{aligned}
\sum_{\boldsymbol{s}^\ell \in J} a_{\boldsymbol{s}^\ell} &\le \frac{|J|}{2^{k'\ell}} \sum_{\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}} a_{\boldsymbol{s}^\ell} + K_2 \sqrt{\frac{|J|}{2^{k'\ell}} \sum_{\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}} a_{\boldsymbol{s}^\ell}^2 \cdot \log m} + K_2 \max_{\boldsymbol{s}^\ell \in \{-1,1\}^{k'\ell}} |a_{\boldsymbol{s}^\ell}| \cdot \log m \\
&\le K_3 \frac{|J|}{2^{k'\ell}} \left(\frac{mp}{1-\alpha}\right)^{\ell-1} + K_3 \frac{(mp)^\ell}{2^{k'\ell}} \sqrt{|J| \cdot \log m} + K_3 \frac{(mp)^\ell}{2^{k'\ell}} \log m \\
&\le 3K_3 \frac{(mp)^\ell}{2^{k'\ell}} \left(\frac{w}{mp(1-\alpha)^{\ell-1}} + \sqrt{w \log m}\right),
\end{aligned} \tag{H.15}$$

with probability at least $1 - m^{-D-3}$ for sufficiently large constants K_2, K_3 depending on K_1, and D. Similarly, we obtain

$$\sum_{\boldsymbol{s}^\ell \in J} a'_{\boldsymbol{s}^\ell} \le 3K_3 \frac{(mp)^\ell}{2^{k'\ell}} \left(\frac{w}{mp(1-\alpha)^{\ell-1}} + \sqrt{w \log m}\right). \tag{H.16}$$

Hence, (H.2) holds from (H.15) and (H.16). By applying a union bound over all distinct vertex pairs $i, i' \in C_k$, the proof is completed. □

Lemma H.4 (Similar to Lemma 6.2 in (Mao et al., 2021a)). *For any constants $D, \delta > 0$, there exists $c \in (0, 1/2)$ and R, Q_1, Q_2, K, K', m_0, which depend on D, δ, with the properties below. Consider a random subset J uniformly drawn from $\{-1,1\}^{k'\ell}$ with cardinality $2w$, where w is an integer satisfying $2w > \log m$. Suppose that*

$$m \ge m_0, \quad (\log m)^{1+\delta} \le mp/(1-\alpha) \le m^{1/20}, \quad mq \ge Q_1 k'^2 \ell^2, \quad k' \le Q_2 \log mp, \quad \ell \le \frac{\log m}{R \log mp}. \tag{H.17}$$

Then, there exists a subset $\mathcal{I} \subset C_k$ with $|\mathcal{I}| \ge m - m^{1-c}$ such that the event $\mathcal{H}$ (or condition $(D3)$) holds for every pair $i, i' \in \mathcal{I}$, where $i \ne i'$, with probability at least $1 - m^{-D}$.

Proof. To apply Lemma H.3, we first show the existence of a subset $\mathcal{I} \subset C_k$ with $|\mathcal{I}| \ge m - m^{1-c}$ such that conditions $(E1), (E2)$ and $(E3)$ are satisfied for all pairs of $i, i' \in \mathcal{I}$, where $i \ne i'$.

- By Proposition F.6, there exists a subset $\mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)$ such that $\left|\mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)\right| \ge m - m^{1-c}$ for $c \in (0, 0.5)$ with probability at least $1 - m^{-D-2}$. Since $G_0\left(\mathcal{B}_{G_0(C_k)}(i, 3\ell)\right)$ and $G_0\left(\mathcal{B}_{G_0(C_k)}(i', 3\ell)\right)$ form trees for any $i, i' \in \mathrm{Typ}_{G_0(C_k)}(3\ell, \epsilon)$, the condition $(E3)$ holds with probability at least $1 - m^{-D-2}$ for K_1 depending on D by Lemma F.2. (Since 3ℓ is greater than or equal to 3, in order to apply Proposition F.6, we have set $mp/(1-\alpha) \le m^{1/20}$.)
- The condition $(E1)$ is directly satisfied since $G_0\left(\mathcal{B}_{G_0(C_k)}(i, 3\ell)\right)$ and $G_0\left(\mathcal{B}_{G_0(C_k)}(i', 3\ell)\right)$ form trees for any $i, i' \in \mathcal{I}$.
- By Lemma F.3, the condition $(E2)$ holds with probability at least $1 - m^{-D-2}$ for any $i, i' \in \mathcal{I}$.

Thus, by applying Lemma H.3, we can have (H.1) and (H.2) for all pairs of $i, i' \in \mathcal{I}$, where $i \ne i'$, with probability at least $1 - 2m^{-D-1}$. The verification of (H.5) follows a similar approach as presented in Lemma 6.2 of (Mao et al., 2021a); thus, we omit the detailed explanation. □

I. Technical Tools

Theorem I.1 (Tail bounds on binomial distribution (Okamoto, 1959)). *Let $X \sim \mathrm{Bin}(n, p)$. It holds that*

$$\mathbb{P}\{X \le nt\} \le \exp\left(-n(\sqrt{p} - \sqrt{t})^2\right), \quad \forall 0 \le t \le p$$
$$\mathbb{P}\{X \ge nt\} \le \exp\left(-2n(\sqrt{t} - \sqrt{p})^2\right), \quad \forall p \le t \le 1$$

Lemma I.2 (Lemma 22 in (Mao et al., 2021b)). *Let $X \sim \mathrm{Bin}(n,p)$ with real number $p \in (0,1)$. For any $t, r > 0$, it holds that*

$$\mathbb{P}\{X \in [t, t+r]\} \leq \frac{Cr}{\sqrt{np(1-p)}}$$

with a universal constant $C > 0$.

Lemma I.3 (Lemma 23 in (Mao et al., 2021b)). *Let $X \sim \mathrm{Bin}(n,p)$ with real number $p \in (0,1)$. If $np(1-p) \geq C$ then for any $r \geq 2$, it holds that*

$$\mathbb{P}\{X > np - r\} \geq \frac{1}{2} + c\left(\frac{r}{\sqrt{np(1-p)}} \wedge 1\right), \quad \mathbb{P}\{X < np + r\} \geq \frac{1}{2} + c\left(\frac{r}{\sqrt{np(1-p)}} \wedge 1\right)$$

with universal constants $C, c > 0$.

Lemma I.4 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$. Define $S_n := X_1 + \ldots + X_n$. Then, for all $t > 0$, we have*

$$\begin{aligned} \mathbb{P}\left\{S_n - E[S_n] \geq t\right\} &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \\ \mathbb{P}\left\{|S_n - E[S_n]| \geq t\right\} &\leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \end{aligned} \tag{I.1}$$

Lemma I.5 (Bernstein's inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that $a \leq X_i \leq b$. Define $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right]$. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2/2}{\sigma^2 n + (b-a)t/3}\right)$$

Lemma I.6 (Bernstein's inequality for sampling without replacement). *Let $\{x_1, \ldots, x_N\}$ be given for a finite positive integer N. For a positive integer $n < N$, let $X_1, \ldots, X_n$ be drawn from $\{x_1, \ldots, x_N\}$ without replacement. Define*

$$a := \min_{i \in [N]} x_i, \quad b := \max_{i \in [N]} x_i, \quad \bar{x} := \frac{1}{N}\sum_{i=1}^N x_i, \quad \sigma^2 := \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^2.$$

Then, for any $t > 0$, we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i - n\bar{x}\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2/2}{\sigma^2 n + (b-a)t/3}\right).$$

Lemma I.7 (Lemma 13 in (Mao et al., 2021b)). *Fix $k \in \mathbb{N}^+, p \in (0,1)$ and $\alpha \in [0, 1-p]$. Let $\Gamma_0 \sim \mathcal{G}\left(k, \frac{p}{1-\alpha}\right)$. We can get two subgraphs Γ and Γ' by independently subsampling every edge of Γ_0 with probability $1-\alpha$. Fix subsets $J, J' \subset [k]$ and a vertex $i \in [k] \setminus (J \cup J')$. For any $t > 0$, we have that with probability at least $1 - 6\exp(-t) - 2\exp\left(\frac{-p|J \cap J'|}{3(1-\alpha)}\right)$,*

$$\left||\mathcal{N}_\Gamma(i;J)| - p|J| - |\mathcal{N}_{\Gamma'}(i;J')| + p|J'|\right| \leq 4\left(t + \sqrt{t\alpha p\,|J \cap J'|} + \sqrt{tp\,|J \triangle J'|}\right)$$

where $\mathcal{N}_\Gamma(i;J)$ is the set of neighbors of i within $J \subset [n]$ in a graph Γ.

Lemma I.8 (Lemma 5.5 in (Mao et al., 2021a)). *Given a constant $S > 0$ and an even integer $k \in \mathbb{N}$, let Ω and Ω' be two finite sets that can be expressed as $\Omega = \bigcup_{i=1}^k \Omega_i$ and $\Omega' = \bigcup_{i=1}^k \Omega'_i$ where $|\Omega'_i| \leq S/k$ for all $i \in [k]$. Additionally, let $w \in \{2, 3, \ldots, k/2\}$, and I be a random subset uniformly drawn from $[k]$ with a cardinality of $2w$. Then, for any given values of $L \geq 1$ and $\rho \in (0, 1/4)$, where ρw is an integer, we have that*

$$\mathbb{P}\left\{\left|\left\{i \in I : \exists j \in I \backslash \{i\} \text{ s.t. } \left|\Omega_i \cap \Omega'_j\right| \geq LS/k^2\right\}\right| \geq 2\rho w\right\} \leq \left(\frac{8w^3}{L}\right)^{\rho w}.$$

Lemma I.9 (Hoeffding's inequality with truncation (Lemma 5.10 in (Mao et al., 2021a))). *Let $X_1, \dots, X_N$ be independent random variables such that $|\mathbb{E}[X_i]| \le \tau$ for $\tau > 0$, and that*

$$\mathbb{P}\{|X_i - \mathbb{E}[X_i]| \ge t\} \le 2\exp\left(\frac{-ct^2}{1+t}\right), \quad \forall t > 0,$$

for a constant $c > 0$ for each $i \in [N]$. Then there exists a constant $C > 0$ depending only on c such that, for any $\epsilon \in (0, 0.1)$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N}\left(X_i^2 - \mathbb{E}\left[X_i^2\right]\right)\right| \ge C\log(N/\epsilon)\sqrt{N\log(1/\epsilon)} + C\tau(\sqrt{N\log(1/\epsilon)} + \log(1/\epsilon))\right\} \le \epsilon.$$

Lemma I.10. *Let n be a positive integer and x be a positive real number. If $nx \le 1$, then $(1+x)^n \le 1 + 2nx$.*

Proof. For $k = 1, \dots, n-1$, we have

$$\binom{n}{k+1}x^{k+1} = \binom{n}{k}x^k\frac{(n-k)x}{k+1} \le \frac{1}{2}\binom{n}{k}x^k \tag{I.2}$$

since $nx \le 1$. Thus, we get

$$(1+x)^n = 1 + \sum_{k=1}^{n}\binom{n}{k}x^k \le 1 + 2nx. \tag{I.3}$$

□

J. Refinement Matching Algorithm

The refinement matching algorithm from Mao et al. (2021a) is summarized below.

Algorithm 4 RefinementMatching (Mao et al., 2021a)

Input: two graphs Γ and Γ' on $[n]$, a permutation $\tilde{\pi} : [n] \to [n]$, and a parameter $\epsilon > 0$
Output: a permutation $\hat{\pi} : [n] \to [n]$
$\pi_0 \leftarrow \tilde{\pi}$
for $t = 1, \dots, \lceil \log_2 n \rceil$ **do**
  for $i = 1, \dots, n$ **do**
    if there is a vertex $i' \in [n]$ such that **then**
      - $|\pi_{t-1}^{-1}(\mathcal{N}_\Gamma(i)) \cap \mathcal{N}_{\Gamma'}(j')| \ge \epsilon^2 pn/512$ for all $j' \in [n]\backslash\{i'\}$
      - $|\pi_{t-1}^{-1}(\mathcal{N}_\Gamma(i)) \cap \mathcal{N}_{\Gamma'}(j')| \ge \epsilon^2 pn/512$ for all $j' \in [n]\backslash\{i'\}$
      - $|\pi_{t-1}^{-1}(\mathcal{N}_\Gamma(j)) \cap \mathcal{N}_{\Gamma'}(i')| \ge \epsilon^2 pn/512$ for all $j \in [n]\backslash\{i\}$
    end if
  end for
  extend π_t to a permutation on $[n]$ in an arbitrary manner
end for
$\hat{\pi} \leftarrow \pi_{\lceil \log_2 n \rceil}$
return $\hat{\pi}$