Random Graph Matching at Otter's Threshold via Counting Chandeliers

Cheng Mao
Georgia Institute of Technology
Atlanta, GA, USA
cheng.mao@math.gatech.edu

Yihong Wu
Yale University
New Haven, CT, USA
yihong.wu@yale.edu

Jiaming Xu
Duke University
Durham, NC, USA
jx77@duke.edu

Sophie H. Yu
Duke University
Durham, NC, USA
haoyang.yu@duke.edu

ABSTRACT

We propose an efficient algorithm for graph matching based on similarity scores constructed from counting a certain family of weighted trees rooted at each vertex. For two Erdős–Rényi graphs $\mathcal{G}(n, q)$ whose edges are correlated through a latent vertex correspondence, we show that this algorithm correctly matches all but a vanishing fraction of the vertices with high probability, provided that $nq \to \infty$ and the edge correlation coefficient ρ satisfies $\rho^2 > \alpha \approx 0.338$, where α is Otter's tree-counting constant. Moreover, this almost exact matching can be made exact under an extra condition that is information-theoretically necessary. This is the first polynomial-time graph matching algorithm that succeeds at an explicit constant correlation and applies to both sparse and dense graphs. In comparison, previous methods either require $\rho = 1 - o(1)$ or are restricted to sparse graphs.

The crux of the algorithm is a carefully curated family of rooted trees called *chandeliers*, which allows effective extraction of the graph correlation from the counts of the same tree while suppressing the undesirable correlation between those of different trees.

CCS CONCEPTS

• **Mathematics of computing** → **Random graphs**; • **Theory of computation** → **Random network models**.

KEYWORDS

graph matching, network alignment, Otter's constant, tree counting, chandeliers

ACM Reference Format:
Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H. Yu. 2023. Random Graph Matching at Otter's Threshold via Counting Chandeliers. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3564246.3585156

STOC '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 978-1-4503-9913-5/23/06…$15.00
https://doi.org/10.1145/3564246.3585156

1 INTRODUCTION

Graph matching (also known as network alignment) refers to the problem of finding the bijection between the vertex sets of the two graphs that maximizes the total number of common edges. When the two graphs are exactly isomorphic to each other, this reduces to the classical graph isomorphism problem, for which the best known algorithm runs in quasi-polynomial time [5]. In general, graph matching is an instance of the *quadratic assignment problem* [10], which is known to be NP-hard to solve or even approximate [29].

Motivated by real-world applications (such as social network de-anonymization [36] and computational biology [43]) as well as the need to understand the average-case computational complexity, a recent line of work is devoted to the study of theory and algorithms for graph matching under statistical models, by assuming the two graphs are randomly generated with correlated edges under a hidden vertex correspondence. A canonical model is the following *correlated Erdős–Rényi graph model* [39].

Definition 1 (Correlated Erdős–Rényi graph model)**.** Let π denote a latent permutation on $[n] \triangleq \{1, \dots, n\}$. We generate two random graphs on the common vertex set $[n]$ with adjacency matrices A and B such that $(A_{ij}, B_{\pi(i)\pi(j)})$ are i.i.d. pairs of Bernoulli random variables with mean $q \in [0, 1]$ and correlation coefficient ρ for $1 \le i < j \le n$. We write $(A, B) \sim \mathcal{G}(n, q, \rho)$.

Given $(A, B) \sim \mathcal{G}(n, q, \rho)$, our goal is to recover the latent vertex correspondence π. The information-theoretic thresholds for both exact and partial recovery have been derived [12, 13, 17, 22, 25, 45] and various efficient matching algorithms have been developed with performance guarantees [16, 18–21, 23, 30, 31]. Despite these exciting progresses, most existing efficient algorithms require the two graphs to be almost perfectly correlated; as such, the problem of polynomial-time recovery with a constant correlation remains largely unresolved except for sufficiently sparse graphs. Specifically, if the correlation ρ is an (unspecified) constant sufficiently close to 1, exact recovery is achievable in polynomial time for graphs whose average degrees satisfy $(1 + \epsilon) \log n \le nq \le n^{\frac{1}{\Theta(\log \log n)}}$ [31], while partial recovery is achievable for sparse graphs with $nq = O(1)$ [21, 23]. For dense graphs, the best known result for polynomial-time recovery requires $\rho \ge 1 - (\log \log(n))^{-C}$ for some constant $C > 0$ [30]. The current paper significantly advances the state of the art by establishing the following results.

Theorem. Assume that $0 < q \le 1/2$ and

$$\rho^2 > \alpha \approx 0.338,$$

where

$$\alpha = \lim_{K\to\infty} \frac{K}{\log(\text{number of unlabeled trees with } K \text{ edges})}$$

is Otter's tree-counting constant [38]. Given a pair of correlated Erdős–Rényi graphs $(A, B) \sim \mathcal{G}(n, q, \rho)$, the following holds:

- *(Exact recovery)* If $\rho > 0$ and $nq(q + \rho(1-q)) \ge (1+\epsilon)\log n$ for any constant $\epsilon > 0$,[1] there is a polynomial-time algorithm that recovers π exactly with high probability.
- *(Almost exact recovery)* If $nq = \omega(1)$, there is a polynomial-time algorithm that outputs a subset $I \subset [n]$ and a map $\hat{\pi} : I \to [n]$ such that $\hat{\pi} = \pi|_I$ and $|I| = (1 - o(1))n$ with high probability.
- *(Partial recovery)* For any constant $\delta \in (0, 1)$, there is a constant $C(\rho, \delta) > 0$ depending only on ρ and δ such that if $nq \ge C(\rho, \delta)$, the above I and $\hat{\pi}$ satisfy that $\hat{\pi} = \pi|_I$ with high probability and $\mathbb{E}[|I|] \ge (1 - \delta)n$.

The above theorem identifies an explicit threshold $\rho^2 > \alpha$ that allows polynomial-time graph matching for both sparse and dense graphs. In certain regimes, the condition for exact recovery in this result is in fact *optimal*, matching the information-theoretic threshold identified in [13, 45] (see Remark 2 and Figure 2 for a detailed discussion). Here we further assume $\rho > 0$ for exact recovery as the current seeded matching algorithms for boosting from almost exact to exact recovery require a positive correlation.

In passing, we remark that after the initial posting of the present paper, [24] proves that a different algorithm proposed earlier in [23, 40] achieves partial recovery (correctly matching $\Omega(n)$ vertices with $o(n)$ errors with high probability) under the same condition of $\rho > \sqrt{\alpha}$. Their algorithm relies on the tree structure of local neighborhoods and thus is restricted to sparse graphs with $nq = O(1)$. Moreover, their results do not provide exact or almost exact recovery.

1.1 Key Challenges and Algorithmic Innovations

A principled approach to graph matching is the following three-step procedure:

(1) *Signature embedding*: Associate to each vertex i in A a *signature* s_i and to each vertex j in B a signature t_j.
(2) *Similarity scoring*: Compute the similarity score Φ_{ij} based on s_i and t_j using a certain similarity measure on the signature space.
(3) *Linear assignment*: Solve max-weight bipartite matching with weights Φ_{ij} either exactly or approximately (e.g., greedy algorithm).

In this way, we reduce the problem from the NP-hard quadratic assignment to the tractable linear assignment. Clearly, the key to this approach is the construction of the similarity scores.

Many existing algorithms for graph matching largely follow this paradigm using similarity scores based on neighborhood statistics [8, 15, 16, 18, 30], spectral methods [20, 43, 44], or convex relaxations [1, 28, 47]. In terms of theoretical guarantees, these methods either require extremely high correlation or are tailored to sparse graphs. Note that two ρ-correlated Erdős–Rényi graphs differ by $\Theta(1-\rho)$ fraction of edges. Thus, to succeed at a constant ρ bounded away from 1, the similarity scores need to be robust to perturbing a constant fraction of edges. All existing algorithms [21, 23, 31] achieving this goal crucially exploit the tree structure of local neighborhoods and are thus restricted to sparse graphs. On the other hand, algorithms that apply to both sparse and dense graphs [18, 20, 30] so far can only tolerate a vanishing fraction of edge perturbation and thus all require $\rho = 1 - o(1)$.

The major algorithmic innovation of this work is a new construction based on *subgraph counts*. Specifically, the signature assigned to a node i is a vector indexed by a family of non-isomorphic subgraphs, where each entry records the total number of subgraphs rooted at i that appear in the graph weighted by the centered adjacency matrix, known as the signed graph count [9] (cf. (1) and (2) for the formal definition). The similarity score for each pair of vertices is the weighted inner product between their signatures. The key to executing this strategy is a carefully curated family of trees called *chandeliers*, which, as we explain next, allows one to extract the graph correlation from the counts of the same tree while suppressing the undesirable correlation between those of different trees. This leads to a robust construction of signatures that can withstand perturbing a constant fraction of edges, without relying on the locally tree-like property that limits the previous methods to sparse graphs.

Counting subgraphs is a popular method for network analysis in both theory [9, 35] and practice [2, 34, 42]. We refer to [32, Sec. 2.4] for a comprehensive overview of hypothesis testing and estimation based on subgraph counting for networks with latent structures. Notably, most of these previous works focus on counting cycles. However, here in order to succeed at a constant ρ, we need to count a sufficiently rich class of subgraphs (whose cardinality grows at least exponentially with the number of edges)[2] and cycles clearly fall short of this basic requirement. A much richer family of *strictly balanced, asymmetric* subgraphs is considered in [6], where the edge density of the subgraphs is carefully chosen so that typically they co-occur in both graphs at most once. Hence, by searching for such rare subgraphs, dubbed "black swans", one can match the corresponding vertices. Although this method succeeds even for vanishing correlation $\rho \ge (\log n)^{-o(1)}$, it has a quasi-polynomial time complexity $n^{\Theta(\log n)}$ due to the exhaustive search of subgraphs of size $\Theta(\log n)$. Moreover, the construction of this special family of subgraphs requires the average degree nq to fall into a very specific range of $[n^{\delta_n}, n^{1/153}] \cup [n^{2/3}, n^{1-\epsilon}]$ for some sequence of positive quantities $\delta_n = o(1)$ and an arbitrarily small constant $\epsilon > 0$, and, in particular, it does not accommodate relatively sparse graphs such as $nq = O(\log n)$.

[1] The condition $nq(q + \rho(1-q)) \ge (1+\epsilon)\log n$ is information-theoretically necessary, for otherwise the intersection graph between A and B (under the vertex correspondence π) contains isolated vertices with high probability and exact recovery is impossible.

[2] A high-level explanation is as follows. For a single subgraph H with N edges, the correlation between the subgraph counts of H rooted at vertex i across A and B – the *signal*, is smaller than their variances by a multiplicative factor of ρ^N. Therefore, to pick up the signal, we need to further average over a family $\mathcal{H}$ of such subgraphs so that $|\mathcal{H}|\rho^{2N} \to \infty$ (cf. (25) for a more detailed explanation).

As opposed to relying on rare subgraphs, our approach is to count a family of unlabeled rooted trees with size $N = \Theta(\log n)$, which are abundant even in very sparse graphs. Moreover, by leveraging the method of *color coding* [3, 4, 26], such trees can be counted approximately but sufficiently accurately in polynomial time. While centering the adjacency matrices and counting signed trees are helpful, there still remains excessive correlation among different trees counts which is hard to control – this is the key difficulty in analyzing signatures based on subgraph counts. To resolve this challenge, we propose to count a special family $\mathcal{T}$ of unlabeled rooted trees, which we call chandeliers; see (1) for the formal definition. As discussed in Section 2.2, the chandelier structure plays a crucial role in curbing the undesired correlation between different tree counts. Moreover, even though chandeliers only occupy a vanishing fraction of all trees, by choosing the parameters appropriately, we can ensure that $|\mathcal{T}| = (1/\alpha + o(1))^N$, which grows almost at the same rate as the entire family of trees.

A similar idea of counting signed but unrooted trees has been applied in [32] for the graph correlation detection problem, i.e., testing whether the two graphs are independent Erdős–Rényi graphs or ρ-correlated through a latent vertex matching chosen uniformly at random. It is shown that the two hypotheses can be distinguished with high probability in polynomial time at the same threshold of $\rho^2 > \alpha$. However, unlike the present paper, averaging over the random permutation dramatically simplifies the analysis of correlations between different tree counts. As a result, it suffices to simply count all trees as opposed to a carefully constructed collection of special trees. We refer to the last two paragraphs in Section 2.2 for a detailed comparison.

1.2 Notation

Given a graph H, let $V(H)$ denote its vertex set and $E(H)$ denote its edge set. Let $v(H) = |V(H)|$ and $e(H) = |E(H)|$. We call $e(H) - v(H)$ the *excess* of the graph H. We denote by $\mathbb{K}_n$ the complete graph with vertex set $[n]$ and edge set $\binom{[n]}{2} \triangleq \{\{u, v\} : u, v \in [n],\ u \neq v\}$. An empty graph is denoted as $\emptyset$, if it does not contain any vertex or edge. A *rooted* graph is a graph in which one vertex has been distinguished as the root. An *isomorphism* between two rooted graphs H and G is a bijection between the vertex sets that preserves both edges and the root, namely, $f : V(H) \to V(G)$ such that the root of H is mapped to that of G and any two vertices u and v are adjacent in H if and only if $f(u)$ and $f(v)$ are adjacent in G. An *automorphism* of a rooted graph is an isomorphism to itself. Let $\mathrm{aut}(H)$ be the number of automorphisms of H. For a rooted tree T and a vertex $a \in V(T)$, let $(T)_a$ denote the subtree of T consisting of all descendants of a and we set $(T)_a = \emptyset$ if $a \notin V(T)$.

For two real numbers x and y, we let $x \vee y \triangleq \max\{x, y\}$ and $x \wedge y \triangleq \min\{x, y\}$. We use standard asymptotic notation: for two positive sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ or $x_n \lesssim y_n$, if $x_n \leq C y_n$ for an absolute constant C and for all n; $x_n = \Omega(y_n)$ or $x_n \gtrsim y_n$, if $y_n = O(x_n)$; $x_n = \Theta(y_n)$ or $x_n \asymp y_n$, if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$; $x_n = o(y_n)$ or $y_n = \omega(x_n)$, if $x_n / y_n \to 0$ as $n \to \infty$.

1.3 Organization

The rest of the paper is organized as follows. In Section 2.1, we first introduce the similarity scores between vertices of the two graphs based on counting signed chandeliers, and then state our main results on the recovery of the latent vertex correspondence for correlated Erdős–Rényi graphs. In Section 2.2, we explain the rationale for focusing on the class of chandeliers. Section 3 provides a statistical analysis of the similarity scores, proving our results on partial and almost exact recovery stated in Theorem 1. In particular, Propositions 2 and 3 are the key ingredients controlling the variance of the similarity scores. In Section 4, we use the method of color coding to approximate the proposed similarity scores in polynomial time, and show that the same statistical guarantees continue to hold for the approximated scores, thereby proving Theorem 2. Finally, in Section 5, we demonstrate how to upgrade an almost exact matching to an exact matching, establishing Theorem 3. Appendix A consists of auxiliary results, and Appendix B discusses a data-driven way to choose a threshold parameter in our algorithm. Due to space constraints, we omit the proofs of Proposition 2-Proposition 6, which can be found in the full version of this paper [33] https://arxiv.org/abs/2209.12313.

2 MAIN RESULTS AND DISCUSSIONS

2.1 Similarity Scores and Statistical Guarantees

We start with some preliminary definitions before specializing to chandeliers. For any weighted adjacency matrix M, node $i \in [n]$, and rooted graph H, define the weighted subgraph count

$$W_{i,H}(M) \triangleq \sum_{S(i) \cong H} M_S\,, \text{ where } M_S \triangleq \prod_{e \in E(S)} M_e\,, \tag{1}$$

and $S(i)$ denotes a subgraph of $\mathbb{K}_n$ rooted at i. (Whenever the context is clear, we also abbreviate $S(i)$ as S.) Note that when M is the adjacency matrix A, $W_{i,H}$ reduces to the usual subgraph count, i.e., the number of subgraphs rooted at i in M that are isomorphic to H. When M is a centered adjacency matrix $\overline{A} \triangleq A - q$, we call $W_{i,H}$ a *signed* subgraph count following [9]. For example (with solid vertex as the root), $W_{i,\bullet\text{-}\circ}(\overline{A}) = d_i - (n-1)q$ and $W_{i,\circ\text{-}\bullet\text{-}\circ}(\overline{A}) = \binom{d_i}{2} - (n-2)d_i q + \binom{n-1}{2} q^2$, where d_i is the degree of i in A.

Next, given a family $\mathcal{H}$ of non-isomorphic rooted graphs H, the *subgraph count signature* of a node i is defined as the vector

$$W_i^{\mathcal{H}}(M) \triangleq \big(W_{i,H}(M)\big)_{H \in \mathcal{H}}\,. \tag{2}$$

Algorithm 1 below describes our proposed method for graph matching based on subgraph count signatures.

At this point Algorithm 1 is a "meta algorithm" and the key to its application is to carefully choose this collection of subgraphs $\mathcal{H}$. Ideally, we would like $\Phi_{ij}^{\mathcal{H}}$ to be maximized at $j = \pi(i)$, at least on average. To this end, we require $H \in \mathcal{H}$ to be *uniquely rooted*, under which we have $\mathbb{E}[\Phi_{ij}^{\mathcal{H}}] \propto \mathbf{1}_{\{\pi(i)=j\}}$ (see Proposition 1).

Definition 2 (Uniquely rooted graph)**.** We say that a graph H rooted at i is uniquely rooted, if $H(i)$ is non-isomorphic to $H(v)$ for any vertex $v \neq i$ in $V(H)$.

However, the uniquely rooted property is far from enough. In order for the signature $\Phi_{ij}^{\mathcal{H}}$ to distinguish whether $j = \pi(i)$ or not, we need to ensure that the fluctuation of $\Phi_{ij}^{\mathcal{H}}$ does not overwhelm

[3]Note that in (3) the coefficient $\mathrm{aut}(H)$ accounts for the symmetry of H and compensates for the fact that the number of copies of H in the complete graph $\mathbb{K}_n$ is inversely proportional to $\mathrm{aut}(H)$. This simplifies the first moment calculation in Proposition 1.

Algorithm 1 Graph Matching by Counting Signed Graphs

1: **Input:** Adjacency matrices A and B on n vertices, a family $\mathcal{H}$ of non-isomorphic rooted graphs, and a threshold $\tau > 0$.
2: **Output:** A mapping $\hat{\pi} : I \to [n]$.
3: For each pair of node i in A and node j in B, compute their similarity score as the *weighted*[3] inner product between their subgraph count signatures:

$$\Phi_{ij}^{\mathcal{H}} \triangleq \left\langle W_i^{\mathcal{H}}(\overline{A}), W_j^{\mathcal{H}}(\overline{B}) \right\rangle \triangleq \sum_{H \in \mathcal{H}} \mathrm{aut}(H)\, W_{i,H}(\overline{A})\, W_{j,H}(\overline{B}), \tag{3}$$

where $\overline{A} = A - q$ and $\overline{B} = B - q$ are the centered adjacency matrices.
4: For each $i \in [n]$, if there exists a unique $j \in [n]$ such that $\Phi_{ij}^{\mathcal{H}} \geq \tau$, let $\hat{\pi}(i) = j$ and include i in set I.

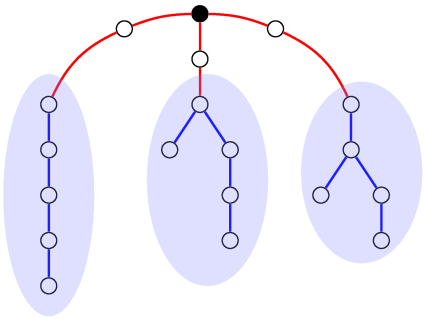

Figure 1: A chandelier with $L = 3$, $M = 2$, $K = 4$, rooted at the solid vertex. The wires are shown in red, and the bulbs in blue. In this case $R = 1$ since each bulb has no non-trivial automorphism (as rooted graphs).

the mean $\mathbb{E}[\Phi_{ii}^{\mathcal{H}}]$ for all $j \in [n]$. In particular, we need $\mathrm{Var}[\Phi_{ij}^{\mathcal{H}}]$ to be much smaller than $(\mathbb{E}[\Phi_{ii}^{\mathcal{H}}])^2$. This turns out to be extremely challenging to show and calls for a rather delicate choice of $\mathcal{H}$. To this end, we construct a special family of trees $\mathcal{T}$, which we call *chandeliers* (see Figure 1 for an illustration).

Definition 3 (Chandelier). An (L, M, K, R)-*chandelier* is a rooted tree with L branches, each of which consists of a path with M edges (which we call an M-*wire*) followed by a rooted tree with K edges (which we call a K-*bulb*); the K-bulbs are non-isomorphic to each other and each of them has at most R automorphisms.

For any chandelier H, let $\mathcal{K}(H)$ denote its set of bulbs. Since all bulbs are *non-isomorphic* to each other, we have

$$\mathrm{aut}(H) = \prod_{\mathcal{B} \in \mathcal{K}(H)} \mathrm{aut}(\mathcal{B}), \tag{4}$$

which is a special case of the classical recursive formula for the number of automorphisms of rooted trees [27]. Moreover, when $L \geq 2$, the root of H is the unique vertex incident to L branches each having $M + K$ edges. As a result, each chandelier is uniquely rooted.

Let $\mathcal{T}$ denote the family of non-isomorphic (L, M, K, R)- chandeliers. Then

$$|\mathcal{T}| = \binom{|\mathcal{J}|}{L}, \tag{5}$$

where $\mathcal{J} \equiv \mathcal{J}(K, R)$ denotes the collection of unlabeled rooted trees having K edges and at most R automorphisms. Counting unlabeled trees with a prescribed number of automorphisms has been well studied in the literature:

- *All trees*: As mentioned earlier in Section 1, a classical result in enumerative combinatorics is that the total number of unlabeled trees with K edges satisfies as $K \to \infty$ [38, 41]

$$|\mathcal{J}(K)| \equiv |\mathcal{J}(K, \infty)| = (\alpha + o(1))^{-K}, \tag{6}$$

where $\alpha \approx 0.338$ is Otter's constant.
- *Typical trees*: The recent result [37] implies that the majority of the trees have $e^{\Theta(K)}$ automorphisms.[4] In other words, for some absolute constant C,

$$|\mathcal{J}(K, \exp(CK))| = (\alpha + o(1))^{-K}. \tag{7}$$

It turns out that to bound the fluctuation of the similarity score it is more advantageous if the bulbs do not have too much symmetry. Thanks to (7) and in view of (4)–(5), by choosing $R = \exp(CK)$ we can ensure that $|\mathcal{T}| = (\alpha + o(1))^{-N}$ has maximal growth while keeping $\mathrm{aut}(H)$ for each $H \in \mathcal{T}$ relatively small.

In the rest of the paper, we will apply the similarity score $\Phi_{ij} \equiv \Phi_{ij}^{\mathcal{T}}$ in (3) to the collection $\mathcal{T}$ of chandeliers with carefully chosen parameters. Crucially, by exploiting the structure of chandeliers, we show:

- For true pairs $j = \pi(i)$,

$$\mathbb{E}[\Phi_{i\pi(i)}] = \mu, \quad \mathrm{Var}(\Phi_{i\pi(i)}) = o(\mu^2),$$

where

$$\mu \triangleq |\mathcal{T}|(\rho\sigma^2)^N \frac{(n-1)!}{(n-N-1)!}, \quad \sigma^2 \triangleq q(1-q). \tag{8}$$

- For fake pairs $j \neq \pi(i)$,

$$\mathbb{E}[\Phi_{ij}] = 0, \quad \mathrm{Var}(\Phi_{ij}) = o\left(\frac{\mu^2}{n^2}\right).$$

This immediately implies that by running a greedy matching with weights Φ_{ij} (or simply thresholding Φ_{ij}), we can match all but a vanishing fraction of vertices correctly with high probability. This is made precise by the following theorem.

Throughout this paper, we assume without loss of generality[5] that $q \leq 1/2$.

Theorem 1 (Partial and almost exact recovery). *There exist absolute constants $C_1, \ldots, C_4 > 0$ such that the following holds. Suppose*

$$\rho^2 \geq \alpha + \epsilon, \tag{9}$$

[4] Indeed, (7) is an immediate corollary of the following asymptotic normality result in [37, Theorem 2]: $\frac{1}{\sqrt{K}}(\log \mathrm{aut}(H_K) - \mu K) \xrightarrow{K \to \infty} N(0, \sigma^2)$, where H_K is a uniform random unlabeled tree with K edges (known as the Pólya tree of order $K+1$), and $\mu \approx 0.137$ and $\sigma^2 \approx 0.197$ are absolute constants.
[5] If $q > 1/2$, we can consider the complement graphs of A and B, which are correlated Erdős–Rényi graphs with parameter $(n, 1-q, \rho)$. In addition, it is not hard to see that the similarity scores Φ_{ij} remain unchanged.

where ϵ is an arbitrarily small constant. Choose $K, L, M, R \in \mathbb{N}$ such that $N = (K+M)L$ is even[6],

$$L = \frac{C_1}{\epsilon}, \quad K = C_2 \log n, \quad M = \frac{C_3 K}{\log(nq)}, \quad R = \exp(C_4 K) \ . \tag{10}$$

Fix any constant $0 < c < 1$ and let μ be given in (8). Let $\hat{\pi} : I \to [n]$ denote the output of Algorithm 1 applied to the collection $\mathcal{T}$ of (L, M, K, R)-chandeliers and threshold $\tau = c\mu$. Then $\hat{\pi} = \pi|_I$ with probability $1 - o(1)$. Moreover,

- *If $nq = \omega(1)$, then $|I| = (1 - o(1))n$ with probability $1 - o(1)$.*
- *For any constant $\delta \in (0, 1)$, there exists a positive constant $C(\epsilon, \delta)$ depending only on ϵ and δ, such that if $nq \geq C(\epsilon, \delta)$, then $\mathbb{E}[|I|] \geq (1 - \delta)n$.*

Remark 1 (Adapting to unknown parameters)**.** Note that the choice of M and τ in (10) assumes the knowledge of q and ρ. The edge probability q can be easily estimated by the empirical graph density of A and B. Moreover, the threshold τ can be specified in a data-driven manner (cf. Appendix B).

From a computational perspective, naïve evaluation of $W_{i,H}(\overline{A})$ by exhaustive search for each H with N edges takes $n^{\Theta(N)}$ time which is super-polynomial when $N = \omega(1)$. To resolve this computational issue, in Section 4, we give a polynomial-time algorithm (Algorithm 2) that computes an approximation $\widetilde{\Phi}_{ij}$ for Φ_{ij} using the strategy of *color coding* as done in [32]. The following result shows that the approximated similarity score $\widetilde{\Phi}_{ij}$ enjoys the same statistical guarantee under the same condition (9) as Theorem 1.

Theorem 2. *Theorem 1 continues to hold with $\widetilde{\Phi}_{ij}$ in place of Φ_{ij}. Moreover, $\{\widetilde{\Phi}_{ij}\}_{i,j\in[n]}$ can be computed in $O(n^C)$ for some constant $C \equiv C(\epsilon)$ depending only on ϵ.*

Theorem 2 shows that our matching algorithm achieves the almost exact recovery in polynomial time when $nq = \omega(1)$ and $\rho^2 \geq \alpha + \epsilon$. In comparison, the almost exact recovery is information-theoretically possible if and only if $nq\rho = \omega(1)$, when $\rho > 0$ and $q = n^{-1/2-\Omega(1)}$ [14, 45].

Moreover, under an extra condition that is information- theoretically necessary, we can upgrade the almost exact recovery to exact recovery in polynomial time. The main idea is to use the partial matching $\hat{\pi}|_I$ correctly identified by Algorithm 1 as *seeds* and apply a seeded matching algorithm (which is similar to percolation-based matching in [6, 46]) to extend it to a full matching. For this purpose we assume $\rho > 0$ as the current seeded matching algorithm requires positive correlation.

Theorem 3 (Exact recovery)**.** *Suppose*

$$nq\,(q + \rho(1-q)) \geq (1+\epsilon)\log n, \quad \rho \geq \sqrt{\alpha + \epsilon} \tag{11}$$

for some arbitrarily small constant ϵ. Then a seeded matching algorithm (see Algorithm 3 in Section 5) with input $\hat{\pi}$ outputs $\widetilde{\pi} = \pi$ in $O(n^3 q^2)$ time with probability $1 - o(1)$.

Remark 2 (Comparison to the exact recovery threshold)**.** It is instructive to compare the performance guarantee (11) of our polynomial-time algorithm with the information-theoretic threshold of exact recovery derived in [45] for positive correlation, that is,

$$\rho \geq (1+\epsilon)\left(2\sqrt{\frac{\log n}{n}} + \frac{\log n}{nq}\right). \tag{12}$$

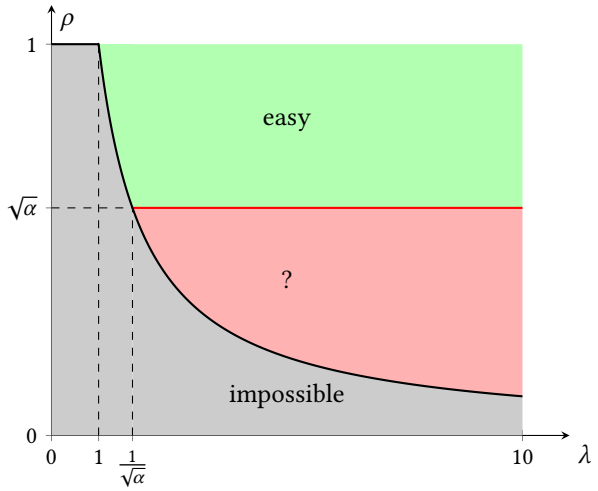

Figure 2: The phase diagram for exact recovery in the logarithmic degree regime, where $nq = \lambda \log n$ for a fixed constant $\lambda > 0$. The impossible and easy regime are given by $\rho < \min\{1, 1/\lambda\}$ and $\rho > \max\{\sqrt{\alpha}, 1/\lambda\}$, respectively. No polynomial-time algorithm is known to achieve exact recovery in the red regime.

Assuming $nq = \lambda \log n$ for a fixed constant λ, (11) simplifies to $\rho > \max\{1/\lambda, \sqrt{\alpha}\}$, while (12) is reduced to $\rho > 1/\lambda$; see Figure 2 for an illustration. Observe that when $\lambda < 1/\sqrt{\alpha}$, the condition (11) for exact recovery matches (12) and hence our polynomial-time matching algorithm is information-theoretically optimal. If $\lambda > 1/\sqrt{\alpha}$, there exists a gap, between (11) and (12), depicted as the red regime in Figure 2. It is an open problem whether exact recovery is attainable in polynomial time in the red regime when $\rho < \sqrt{\alpha}$. So far the only rigorous evidence for hardness is that detection (and hence recovery) is computationally hard in the low-degree polynomial framework[7] when $\rho \leq 1/\mathrm{polylog}(n)$ [32].

2.2 On the Choice of Chandeliers

The key to the success of our matching algorithm is to leverage the correlation of subgraph counts in the two graphs A and B as much as possible, while suppressing the undesired correlation between different subgraph counts. In this subsection, we explain why restricting to the special family of chandeliers is crucial, as well as some basic guidelines on the choice of its parameters. Assume for convenience that $\pi = \mathrm{id}$.

First of all, we require the expected similarity score $\mathbb{E}[\Phi_{ij}]$ to be zero except for $i = j$. As discussed in the previous subsection, this is guaranteed by the uniquely rooted property of each chandelier in $\mathcal{T}$. Further, to distinguish a true pair (i, i) from fake pairs (i, j), we

[6] For simplicity, we assume N is even so that $\mu \geq 0$ even when $\rho < 0$. To lighten the notation, we do not explicitly round each parameter in (10) to integers as this only changes constant factors; see (26) for a more general condition.

[7] Specifically, it is shown in [32] that any test statistic that is a degree-polylog(n) polynomial of (A, B) fails to detect correlation $\rho = 1/\mathrm{polylog}(n)$.

need $\mathrm{Var}(\Phi_{ij})$ to be much smaller than $\mathbb{E}[\Phi_{ii}]^2$ for any pair (i, j). More precisely, in order to apply a union bound over all fake pairs, we need $\mathrm{Var}(\Phi_{ij})/\mathbb{E}[\Phi_{ii}]^2 = o(1/n^2)$ for all $i \neq j$. Later in (25), we will see that even if all the tree counts were uncorrelated, the variance would always be lower bounded by $\mathrm{Var}(\Phi_{ij})/\mathbb{E}[\Phi_{ii}]^2 = \Omega(|\mathcal{T}|^{-1}\rho^{-2N})$. It follows that our class of chandeliers $\mathcal{T}$ needs to satisfy

$$|\mathcal{T}| = \omega(n^2\rho^{-2N}). \tag{13}$$

By choosing the parameters appropriately, we can ensure that $|\mathcal{T}|$ grows as $(\alpha + o(1))^{-N}$, almost at the same rate as the entire set of unlabelled rooted trees. Therefore, whenever $\rho^2 > \alpha$, (13) holds by choosing $N = \Theta(\log n)$.

To further see the significance of chandeliers on the correlations between subgraph counts, let us expand out the variance of Φ_{ij}:

$$\begin{aligned}\mathrm{Var}[\Phi_{ij}] &= \sum_{H,I\in\mathcal{T}} \mathrm{aut}(H)\mathrm{aut}(I)\cdot \\ &\qquad \mathrm{Cov}\left(W_{i,H}(\overline{A})W_{j,H}(\overline{B}), W_{i,I}(\overline{A})W_{j,I}(\overline{B})\right) \\ &= \sum_{H,I\in\mathcal{T}} \mathrm{aut}(H)\mathrm{aut}(I) \sum_{S_1(i),S_2(j)\cong H} \sum_{T_1(i),T_2(j)\cong I} \\ &\qquad \mathrm{Cov}\left(\overline{A}_{S_1}\overline{B}_{S_2}, \overline{A}_{T_1}\overline{B}_{T_2}\right). \end{aligned} \tag{14}$$

Here, S_1 and T_1 are labeled subgraphs of $\mathbb{K}_n$ isomorphic to chandeliers H and I respectively and both rooted at i, and similarly for S_2 and T_2 rooted at j. It turns out that, thanks to centering, $\mathrm{Cov}\left(\overline{A}_{S_1}\overline{B}_{S_2}, \overline{A}_{T_1}\overline{B}_{T_2}\right) = 0$ unless every edge in the union graph $U \triangleq S_1 \cup T_1 \cup S_2 \cup T_2$ appears *at least twice* in the 4-tuple (S_1, T_1, S_2, T_2). Furthermore, each covariance in (14) is upper bounded by $\sigma^{4N}q^{-2N+e(U)}$ (cf. [33, eq(45)]). To proceed, we need to enumerate all possible 4-tuples (S_1, T_1, S_2, T_2) according to the union graph U. Note that the number of different vertex labelings of U (excluding vertices i and j) is simply upper bound by $n^{v(U)-1-1_{\{i\neq j\}}}$. However, there are many configurations for the four chandeliers (S_1, T_1, S_2, T_2) to generate the same unlabeled graph U, which may lead to excessive correlation. The chandelier structure is designed specifically to limit the possible overlapping patterns and reduce the correlations.

To convey some intuitions, let us focus on a true pair (i, i) and consider the simple case where U is a tree and every edge in U appears exactly twice in the 4-tuple. In this case, $e(U) = 2N$ and $v(U) = 2N + 1$. Moreover, U is a chandelier with $2L$ branches, each of which belongs to exactly two out of the four chandeliers (S_1, T_1, S_2, T_2). For example, in Figure 3(a), we show two branches of U, one comes from S_1, S_2 and the other comes from T_1, T_2. Using this specific structure, we can precisely enumerate all possible 4-tuples that generate such a union graph U and bound their contributions to the variance.

Moving from this simple case (referred to as the baseline) to more general cases, the following three observations are crucial for bounding the total variance (although the proof does not exactly follow this classification):

- If bulbs from different branches overlap (Figure 3(b)), this will create cycles and hence increase the excess $e(U)-v(U)$, gaining extra factors of $1/n$ in the variance bound (14) compared to the contribution of the baseline. As a result, although the structure of U is difficult to track, a crude enumeration based on $e(U)$ and $v(U)$ suffices. Next we assume U is a *tree*.
- If two wires completely overlap (Figure 3(c)), both $e(U)$ and $v(U)$ are reduced by M and hence we gain a factor of $(nq)^{-M}$ in the variance bound. On the other hand, the two bulbs can intersect to form an arbitrary tree which has at most $\exp(O(K))$ possibilities up to isomorphism. To ensure $(nq)^{-M}$ dominates $\exp(O(K))$, we need $M \gtrsim K/\log(nq)$.
- If two wires first intersect then branch out (Figure 3(d)), the attached bulbs must be disjoint (otherwise a cycle will ensue), so that each bulb appears in exactly two out of (S_1, T_1, S_2, T_2). It turns out that the worst case occurs when the two wires share a single edge, for which there are at most L^2 possible ways (since each chandelier has L wires). On the other hand, we gain a factor of $(nq)^{-1}$ in the variance bound (14) (cf. Remark 3). Thus to ensure L^2 is dominated by nq^{-1}, we need $L = o(\sqrt{nq})$.

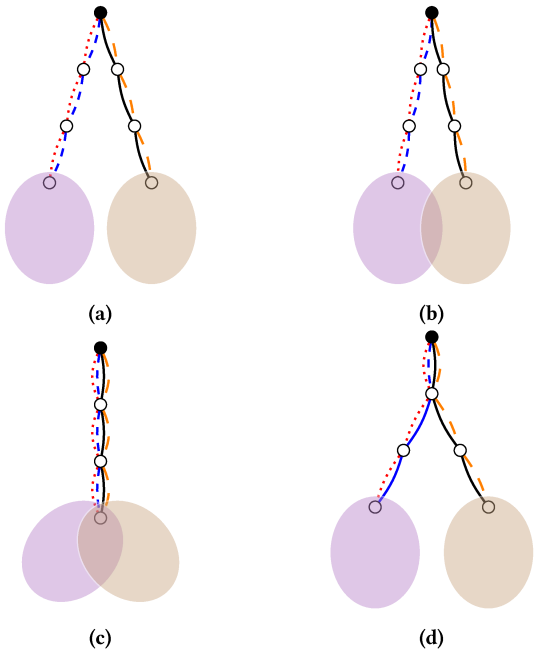

Figure 3: Examples of overlapping patterns of two branches from, say, $S_1 \cap S_2$ and $T_1 \cap T_2$, shown in red/blue and black/orange respectively. The solid vertex is the root i. (a): The two branches overlap only at the root i. (b): The two wires are disjoint and the two dangling bulbs intersect creating cycle(s). (c): The two wires completely overlap and the bulbs can intersect into an arbitrary tree. (d): The two wires overlap in the beginning before branching out and the bulbs are disjoint.

In all, we see that it is critical for chandeliers to be a "thin" tree with only a few long wires, especially when the graphs get sparser. To further reduce the symmetry, we require the bulbs in each chandelier H are all non-isomorphic so that $\mathrm{aut}(H)$ is given by (4), namely $\mathrm{aut}(H) = \prod_{\mathcal{B}\in\mathcal{K}(H)} \mathrm{aut}(\mathcal{B})$, and each $\mathrm{aut}(\mathcal{B})$ is required to be at most $R = \exp(O(K))$.

The method of counting signed trees has been applied to the *detection* problem in the previous work [32]. The goal therein is to decide whether two Erdős–Rényi random graphs are independent or correlated using the test statistic

$$f(A,B) = \sum_{H\in\mathcal{T}'} \mathrm{aut}(H)\, W_H(\overline{A})\, W_H(\overline{B}), \tag{15}$$

where the weighted subgraph count $W_H(W)$ is similarly defined as (1) for unrooted H. Compared to (3), there are three major distinctions: First, the trees in (15) are not rooted and $\mathrm{aut}(\cdot)$ is for unrooted graphs. Second, trees in $\mathcal{T}'$ only have $\Theta(\log n/\log\log n)$ edges, instead of $\Theta(\log n)$ edges required in this paper. This is because for detection, one only needs to achieve a vanishing error, instead of a specific $o(1/n^2)$ error probability for recovery in the current work. Third (and most importantly), $\mathcal{T}'$ contains all trees without special structure, while here we choose $\mathcal{T}$ to be a family of special trees called chandeliers, which, as explained earlier, is crucial for reducing the correlation between different signed tree counts.

In terms of analysis, for the detection problem in [32] the latent permutation is chosen uniformly at random, so one can average the second moment calculation over the random permutation which drastically simplifies the analysis of the tree counting statistic. In contrast, for the recovery problem in the present paper, we need to condition on the realization of the latent permutation. As such, the second moment calculation here is much more challenging combinatorially and involves delicate enumeration procedures that revolve around the chandelier construction. In addition, since the trees in [32] are much smaller with only $\Theta(\frac{\log n}{\log\log n})$ edges, so that many quantities can be bounded very crudely (e.g., $\mathrm{aut}(H)\le v(H)!$); for the current paper since the trees have $\Theta(\log n)$ edges such simple analysis does not suffice.

3 STATISTICAL ANALYSIS OF SIMILARITY SCORES

Throughout the analysis, without loss of generality, we assume $\pi = \mathrm{id}$. First, we compute the first moment of the similarity scores $\Phi_{ij}^{\mathcal{H}}$ for a general collection $\mathcal{H}$ of subgraphs.

Proposition 1. *Let $\mathcal{H}$ be a family of unlabeled uniquely rooted graphs with N edges and $V+1$ vertices. For any $i,j\in[n]$, we have*

$$\mathbb{E}\left[\Phi_{ij}^{\mathcal{H}}\right] = |\mathcal{H}|\left(\rho\sigma^2\right)^N \frac{(n-1)!}{(n-V-1)!}\mathbf{1}_{\{i=j\}}, \tag{16}$$

where $\sigma^2 = q(1-q)$. Moreover, if $V^2 = o(n)$, then we have $\mathbb{E}\left[\Phi_{ij}^{\mathcal{H}}\right] = (1+o(1))|\mathcal{H}|\left(\rho\sigma^2\right)^N n^V$.

Proof. For a rooted graph H with N edges and $V+1$ vertices, the number of copies of H in the complete graph $\mathbb{K}_n$ that are rooted at $i\in[n]$ is

$$\mathrm{sub}_n(H) \equiv \mathrm{sub}(H,\mathbb{K}_n) = \frac{\binom{n-1}{V}V!}{\mathrm{aut}(H)}, \tag{17}$$

where recall that $\mathrm{aut}(H)$ denotes the number of automorphisms of H. For any weighted adjacency matrix M and any subgraph S of $\mathbb{K}_n$, recall that $M_S = \prod_{e\in E(S)} M_e$ as in (1). Then,

$$\begin{aligned}\mathbb{E}\left[W_{i,H}(\overline{A})W_{j,H}(\overline{B})\right] &= \sum_{S(i)\cong H}\sum_{T(j)\cong H}\mathbb{E}\left[\overline{A}_S\overline{B}_T\right]\\ &\overset{(a)}{=} \sum_{S(i)\cong H, S(j)\cong H}\mathbb{E}\left[\overline{A}_S\overline{B}_S\right]\\ &\overset{(b)}{=} \left(\rho\sigma^2\right)^N \mathrm{sub}_n(H)\mathbf{1}_{\{i=j\}},\end{aligned} \tag{18}$$

where (a) is because $\mathbb{E}\left[\overline{A}_S\overline{B}_T\right] = 0$ unless $S=T$ (as unrooted graphs); (b) is because $S(i)\cong H$ and $S(j)\cong H$ imply that $i=j$, thanks to the unique-rootedness of H. By (3),

$$\begin{aligned}\mathbb{E}[\Phi_{ij}^{\mathcal{H}}] &= \sum_{H\in\mathcal{H}}\mathrm{aut}(H)\mathbb{E}[W_{i,H}(\overline{A})W_{j,H}(\overline{B})]\\ &= |\mathcal{H}|\left(\rho\sigma^2\right)^N\binom{n-1}{V}V!\mathbf{1}_{\{i=j\}}.\end{aligned}$$

In view of $\binom{n-1}{V}V! = \frac{(n-1)!}{(n-V-1)!}$, we obtain the desired (16). Finally, since $\left(1-\frac{V}{n}\right)^V \le \frac{(n-1)!}{(n-V-1)!n^V} \le \left(1-\frac{1}{n}\right)^V$ and $V=o(\sqrt{n})$, we have $\frac{(n-1)!}{(n-V-1)!} = (1+o(1))n^V$. □

Next, we bound the variance of the similarity scores $\Phi_{ij}\equiv\Phi_{ij}^{\mathcal{T}}$, where $\mathcal{T}$ is the collection of (K,L,M,R)-chandeliers, for both true pairs $i=j$ and fake pairs $i\ne j$. In the remainder of the paper, let β denote a universal constant such that

$$|\mathcal{J}(K)| \le \beta^K, \quad \forall K\ge 1. \tag{19}$$

Such a β (not to be confused with Otter's constant α) exists thanks to (6).

Proposition 2 (True pairs). *Suppose $q\le\frac{1}{2}$, $L\ge 2$, and*

$$\begin{aligned}&\frac{14L^2}{\rho^{2(K+M)}|\mathcal{J}|}\le\frac{1}{2}, \quad \frac{11R^4(2N)^3(11\beta)^{2(K+M)}}{n}\le\frac{1}{2},\\ &\frac{R^{\frac{4}{M}}(11\beta)^{\frac{4M+4K}{M}}}{nq}\le\frac{1}{2}, \quad \frac{1+2L^2}{\rho^2 nq}\le\frac{1}{2}.\end{aligned} \tag{20}$$

Then, for any $i\in[n]$, we have

$$\frac{\mathrm{Var}\left[\Phi_{ii}\right]}{\mathbb{E}\left[\Phi_{ii}\right]^2} = O\left(\frac{L^2}{\rho^2 nq}+\frac{L^2}{\rho^{2(K+M)}|\mathcal{J}|}\right). \tag{21}$$

Proposition 3 (Fake pairs). *Suppose $q\le\frac{1}{2}$, $L\ge 2$, and*

$$\begin{aligned}&\frac{R^{\frac{2}{M}}(11\beta)^{\frac{4(K+M)}{M}}}{nq}\le\frac{1}{2},\\ &4^{L+3}L^{2L\wedge(4K+2)}(11\beta)^{8(K+M)}R^2(2N+1)^3\le\frac{n}{2}.\end{aligned} \tag{22}$$

Then, for any $i\ne j$, we have

$$\frac{\mathrm{Var}\left[\Phi_{ij}\right]}{\mathbb{E}\left[\Phi_{ii}\right]^2} = O\left(\frac{1}{|\mathcal{T}|\rho^{2N}}\right). \tag{23}$$

The next remark shows that the results in Propositions 2 and 3 are essentially optimal, by identifying which configurations of (S_1,T_1,S_2,T_2) in (14) contribute predominantly to the variance.

Remark 3. The upper bound (21) for true pairs is almost tight. In fact, when $N^2 \ll n$, $q = o(1)$ and $\rho \geq 0$,

$$\frac{\mathrm{Var}\left[\Phi_{ii}\right]}{\mathbb{E}\left[\Phi_{ii}\right]^2} \geq \Omega\left(\frac{L^2}{nq} + \frac{L^2}{\rho^{2(K+M)}|\mathcal{J}|}\right). \tag{24}$$

For the first term in this lower bound, fix any $H, I \in \mathcal{T}$ and consider the special case where $S_1 = S_2 \cong H$, $T_1 = T_2 \cong I$, where S_1 and T_1 only intersect on one edge that connects to i (see Figure 3(d)). Then, we can show that $\mathrm{Cov}\left(\overline{A}_{S_1}\overline{B}_{S_2}, \overline{A}_{T_1}\overline{B}_{T_2}\right) = \Omega(\left(\rho\sigma^2\right)^{2N} q^{-1})$. There are $\Omega(L^2 \mathrm{sub}_n(H)\mathrm{sub}_n(I)n^{-1})$ number of (S_1, T_1, S_2, T_2) that satisfies the above condition. Combining this with (14) and applying Proposition 1, we obtain

$$\begin{aligned}\frac{\mathrm{Var}\left[\Phi_{ii}\right]}{\mathbb{E}\left[\Phi_{ii}\right]^2} &\gtrsim \frac{1}{|\mathcal{T}|^2\left(\rho\sigma^2\right)^{2N} n^{2N}} \\ &\quad \sum_{H,I\in\mathcal{T}} \mathrm{aut}(H)\mathrm{aut}(I)\mathrm{sub}_n(H)\mathrm{sub}_n(I)\left(\rho\sigma^2\right)^{2N} L^2(nq)^{-1} \\ &= \Omega\left(\frac{L^2}{nq}\right),\end{aligned}$$

where the last equality holds because $\mathrm{aut}(H)\mathrm{sub}_n(H) = \Omega(n^N)$ by (17) and $N^2 \ll n$.

For the second term in (24), suppose the chandeliers H and I only share one common bulb $\mathcal{B}$ (i.e., $|\mathcal{K}(H) \cap \mathcal{K}(I)| = 1$). Consider (S_1, T_1, S_2, T_2) such that (i) S_1 (resp. T_1) completely overlaps with S_2 (resp. T_2) except for $\mathcal{B}$ and its attached wire; (ii) S_1 (resp. S_2) only overlaps with T_1 (resp. T_2) on $\mathcal{B}$ and its attached wire. This corresponds to a baseline case as described in Section 2.2. Then, we can show $\mathrm{Cov}\left(\overline{A}_{S_1}\overline{B}_{S_2}, \overline{A}_{T_1}\overline{B}_{T_2}\right) \geq \left(\rho\sigma^2\right)^{2N}\rho^{-2(M+K)}$, and there are $\Omega(\mathrm{sub}_n(H)\mathrm{sub}_n(I))$ number of (S_1, T_1, S_2, T_2) satisfying the above conditions (i) and (ii). Therefore, combining this with (14) and Proposition 1 yields

$$\begin{aligned}\frac{\mathrm{Var}\left[\Phi_{ii}\right]}{\mathbb{E}\left[\Phi_{ii}\right]^2} &\gtrsim \frac{1}{|\mathcal{T}|^2\left(\rho\sigma^2\right)^{2N} n^{2N}} \sum_{H,I\in\mathcal{T}} \mathrm{aut}(H)\mathrm{aut}(I)\mathrm{sub}_n(H)\mathrm{sub}_n(I) \\ &\quad \left(\rho\sigma^2\right)^{2N}\rho^{-2(M+K)}\mathbf{1}_{\{|\mathcal{K}(H)\cap\mathcal{K}(I)|=1\}} \\ &= \frac{\sum_{H,I\in\mathcal{T}}\mathbf{1}_{\{|\mathcal{K}(H)\cap\mathcal{K}(I)|=1\}}}{|\mathcal{T}|^2\rho^{2(M+K)}} \gtrsim \frac{L^2}{\rho^{2(K+M)}|\mathcal{J}|}.\end{aligned}$$

where the last step holds because there are $L\binom{|\mathcal{J}|}{L}\binom{|\mathcal{J}|}{L-1}$ number of pairs of H and I that only share a single bulb.

The upper bound (23) for fake pairs is sharp. In fact, if $N^2 \ll n$, $q \leq 1/2$, and $\rho \geq 0$, for any collection $\mathcal{H}$ of uniquely rooted trees (not just chandeliers) and any fake pair $i \neq j$, we have

$$\frac{\mathrm{Var}[\Phi_{ij}^{\mathcal{H}}]}{\mathbb{E}\left[\Phi_{ii}^{\mathcal{H}}\right]^2} \geq \Omega\left(\frac{1}{|\mathcal{H}|\rho^{2N}}\right). \tag{25}$$

To see this, first note that for any S_1, T_1, S_2, T_2 where $S_1(i), S_2(j) \cong H$ and $T_1(i), T_2(j) \cong I$ with $H, I \in \mathcal{H}$,

$$\begin{aligned}&\mathrm{Cov}\left(W_{i,H}(\overline{A})W_{j,H}(\overline{B}),\, W_{i,I}(\overline{A})W_{j,I}(\overline{B})\right) \\ &= \mathbb{E}\left[W_{i,H}(\overline{A})W_{j,H}(\overline{B})W_{i,I}(\overline{A})W_{j,I}(\overline{B})\right] \geq 0,\end{aligned}$$

where the first equality applies (18) for uniquely rooted trees, and the last inequality holds because $\mathbb{E}\left[\overline{A}_{S_1}\overline{B}_{S_2}\overline{A}_{S_1}\overline{B}_{S_2}\right] \geq 0$ whenever $\rho \geq 0$ (cf. (43) in Lemma 1, Appendix A). Second, consider the special case where $H = I$ and $S_1 = T_1$, $S_2 = T_2$, S_1 and S_2 are vertex-disjoint (i.e., just focus on the diagonal terms in the expansion of the variance (14) and ignore the possible correlations between counts of distinct trees in $\mathcal{H}$), we get

$$\begin{aligned}&\mathrm{Cov}\left(W_{i,H}(\overline{A})W_{j,H}(\overline{B}), W_{i,H}(\overline{A})W_{j,H}(\overline{B})\right) \\ &\geq \sum_{S_1(i)=T_1(i)\cong H}\ \sum_{S_2(j)=T_2(j)\cong H} \\ &\quad \mathbf{1}_{\{S_1 \text{ and } S_2 \text{ are vertex-disjoint}\}}\mathrm{Cov}\left(\overline{A}_{S_1}\overline{B}_{S_2}, \overline{A}_{T_1}\overline{B}_{T_2}\right) \\ &= \sigma^{4N}\sum_{S_1(i)\cong H}\ \sum_{S_2(j)\cong H}\mathbf{1}_{\{S_1 \text{ and } S_2 \text{ are vertex-disjoint}\}} \\ &= \Omega\left(\sigma^{4N}n^{2N}/\mathrm{aut}^2(H)\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left[\Phi_{ij}^{\mathcal{H}}\right] &\geq \sum_{H\in\mathcal{H}}\mathrm{aut}(H)^2 \\ &\quad \mathrm{Cov}\left(W_{i,H}(\overline{A})W_{j,H}(\overline{B}), W_{i,H}(\overline{A})W_{j,H}(\overline{B})\right) \\ &\geq \Omega\left(|\mathcal{H}|\sigma^{4N}n^{2N}\right).\end{aligned}$$

Combining the above with Proposition 1 yields (25).

3.1 Proof of Theorem 1

We aim to prove Theorem 1 under the assumption (9), that is, $\rho^2 \geq \alpha + \epsilon$, and the following more general condition than (10):

$$\begin{aligned}&L \leq \frac{c_1\log n}{\log\log n} \wedge c_6\sqrt{nq}, \quad \frac{c_2}{\log(nq)} \leq \frac{M}{K} \leq \frac{\log\frac{\rho^2}{\alpha}}{2\log\frac{1}{\rho^2}}, \\ &KL \geq \frac{c_3\log n}{\log\frac{\rho^2}{\alpha}}, \quad K + M \leq c_4\log n, \quad R = \exp(c_5K),\end{aligned} \tag{26}$$

for some absolute constants $c_1, \ldots, c_6 > 0$. Indeed, the specific choice of K, L, M, R in (10) satisfies (26) when $nq \geq C(\epsilon)$ for a sufficiently large constant $C(\epsilon)$ that only depends on ϵ.

Next, we verify that (26) with appropriately chosen $(c_1, \ldots, c_6)$ ensures that the condition (20) in Proposition 2 and the condition (22) in Proposition 3 are both satisfied for all sufficiently large n. To start, we note that

$$\frac{M}{K} \leq \frac{\log\frac{\rho^2}{\alpha}}{2\log\frac{1}{\rho^2}} \iff \frac{\rho^{2(K+M)/K}}{\alpha} \geq \sqrt{\frac{\rho^2}{\alpha}}. \tag{27}$$

Moreover, since $R = \exp(c_5K)$, by choosing c_5 to be an appropriate absolute constant and applying (7), we have that for all K large enough,

$$|\mathcal{J}| \geq (\alpha(1 + c_0))^{-K}, \tag{28}$$

where $c_0 > 0$ is an arbitrarily small constant. Combining the last two displayed equation gives that

$$\rho^{2(K+M)}|\mathcal{J}| \geq \left(\frac{\rho^{2(K+M)/K}}{\alpha(1+c_0)}\right)^K \geq \left(\frac{\rho^2}{\alpha}\right)^{K/4}, \tag{29}$$

where the last inequality holds by choosing $c_0 = \rho^2/\alpha - 1 \geq \epsilon/\alpha$. Since $L \leq \frac{c_1 \log n}{\log\log n}$ and $KL \geq \frac{c_3 \log n}{\log(\rho^2/\alpha)}$, $K \geq \frac{c_3 \log\log n}{c_1 \log(\rho^2/\alpha)}$. We deduce from (29) that

$$\rho^{2(K+M)}|\mathcal{J}| \geq (\log n)^{c_3/(4c_1)} \geq \omega(L^2), \tag{30}$$

where the last inequality holds by choosing c_1, c_3 so that $c_3/c_1 > 8$.

Assuming that $K + M \leq c_4 \log n$, $L \leq c_1 \frac{\log n}{\log\log n}$, and $R = \exp(c_5 K)$, by choosing c_4 to be a sufficiently small constant and noting that $N = (K+M)L$, we deduce that

$$\frac{11R^4(2N)^3(11\beta)^{2(K+M)}}{n} \leq \frac{1}{2}.$$

Assuming that $M/K \geq c_2/\log(nq)$, by choosing c_2 to be a sufficiently large constant, we get that

$$\frac{R^{\frac{4}{M}}(11\beta)^{\frac{4M+4K}{M}}}{nq} \leq \frac{1}{2}.$$

Finally, assuming that $L \leq c_6\sqrt{nq}$ and $\rho^2 > \alpha$, by choosing c_6 to be a sufficiently small constant, we conclude that

$$\frac{1+2L^2}{\rho^2 nq} \leq \frac{1}{2}$$

completing the verification of (20).

For (22), under the assumption $L \leq c_1 \frac{\log n}{\log\log n}$, $L^L \leq n^{c_1}$. Thus, under the assumptions that $K + M \leq c_4 \log n$, and $R = \exp(c_5 K)$, by choosing c_1, c_4 to be sufficiently small constants, we get that

$$4^{L+3}L^{2L\wedge(4K+2)}(11\beta)^{8(K+M)}R^2(2N+1)^3 \leq \frac{n}{2},$$

hence the desired(22).

Now we are ready to prove Theorem 1 by applying Propositions 1 and 3. Define

$$F = \{i : |\Phi_{ii} - \mu| > (1-c)\mu\} \supset \{i : \Phi_{ii} < \tau\}, \tag{31}$$

in view of $\tau = c\mu$. Applying Proposition 1, Proposition 3, and Chebyshev's inequality, we get that for any $i \neq j$,

$$\begin{aligned}\mathbb{P}\left\{\Phi_{ij} \geq \tau\right\} &= \mathbb{P}\left\{\Phi_{ij} - \mathbb{E}\left[\Phi_{ij}\right] \geq c\mathbb{E}\left[\Phi_{ii}\right]\right\} \\ &\leq \frac{\mathrm{Var}\left[\Phi_{ij}\right]}{c^2\mathbb{E}\left[\Phi_{ii}\right]^2} = O\left(\frac{1}{|\mathcal{T}|\rho^{2N}}\right).\end{aligned} \tag{32}$$

Note that

$$\begin{aligned}|\mathcal{T}|\rho^{2N} &= \binom{|\mathcal{J}|}{L}\rho^{2N} \geq \left(\frac{|\mathcal{J}|}{L}\right)^L \rho^{2L(K+M)} \\ &\geq \left(\frac{1}{L}\right)^L \left(\frac{\rho^2}{\alpha}\right)^{KL/4} \geq n^{c_3/4 - c_1} = \omega(n^2),\end{aligned} \tag{33}$$

where the second inequality holds due to (29); the last inequality holds due to the assumptions that $L \leq c_1 \log n/\log\log n$ and $KL \geq c_3 \log n/\log(\rho^2/\alpha)$; the last equality holds by choosing $c_3/4 - c_1 > 2$.

Hence, applying union bound together with (32) yields that

$$\mathbb{P}\left\{\exists i \neq j \in [n] : \Phi_{ij} \geq \tau\right\} = o(1). \tag{34}$$

It follows that with probability at least $1 - o(1)$, $\Phi_{ij} < \tau$ for all $i \neq j \in [n]$, which, by our construction of I and $\hat{\pi}$, further implies further implies $I \supset [n] \setminus F$ and $\hat{\pi} = \pi|_I$.

By Chebyshev's inequality and our choice of $\tau = c\mathbb{E}\left[\Phi_{ii}\right] \geq 0$, for any $i \in [n]$,

$$\begin{aligned}\mathbb{P}\left\{|\Phi_{ii} - \mu| > (1-c)\mu\right\} &= \mathbb{P}\left\{|\Phi_{ii} - \mathbb{E}\left[\Phi_{ii}\right]| > (1-c)\mathbb{E}\left[\Phi_{ii}\right]\right\} \\ &\leq \frac{\mathrm{Var}\left[\Phi_{ii}\right]}{(1-c)^2\mathbb{E}\left[\Phi_{ii}\right]^2} \triangleq \gamma,\end{aligned}$$

Applying Proposition 2 yields that

$$\gamma = O\left(\frac{L^2}{nq} + \frac{L^2}{\rho^{2(K+M)}|\mathcal{J}|}\right). \tag{35}$$

It follows that $\mathbb{E}\left[|F|\right] \leq \gamma n$. For any constant $\delta \in (0,1)$, we can choose the constant $C(\epsilon,\delta)$ large enough, so that when $nq \geq C(\epsilon,\delta)$, the assumption $L \leq c_6\sqrt{nq}$ holds for a sufficiently small constant c_6 and consequently $\gamma \leq \delta$. Thus, $\mathbb{E}\left[|I|\right] = n - \mathbb{E}\left[|F|\right] \geq (1-\delta)n$.

If $nq = \omega(1)$, then by choosing $c_6 = o(1)$ we get $\gamma = o(1)$. Therefore, by Markov's inequality,

$$\mathbb{P}\left\{|F| \geq \sqrt{\gamma}n\right\} \leq \sqrt{\gamma} = o(1).$$

It follows that with probability at least $1 - o(1)$, $|F| \leq \sqrt{\gamma}n$ and hence $|I| \geq (1 - \sqrt{\gamma})n = (1 - o(1))n$.

4 APPROXIMATED SIMILARITY SCORES BY COLOR CODING

In this section, following [32], we provide a polynomial-time algorithm to approximately compute the similarity scores $\{\Phi_{ij}\}_{i,j\in[n]}$ in (3) when $\mathcal{T}$ is the family of chandeliers[8] of size $O(\log n)$, using the idea of color coding [2, 3].

Approximate signed rooted subgraph count. Let H be a rooted connected graph with $N+1$ vertices. For each $i \in [n]$, we first approximately count the signed graphs rooted at i that are isomorphic to H. Specifically, given a weighted adjacency matrix M on $[n]$, we generate a random coloring $\mu : [n] \to [N+1]$ that assigns a color to each vertex of M from the color set $[N+1]$ independently and uniformly at random. Given any $V \subset [n]$, let $\chi_\mu(V)$ indicate that $\mu(V)$ is colorful, i.e., $\mu(x) \neq \mu(y)$ for any distinct $x, y \in V$. In particular, if $|V| = N+1$, then $\chi_\mu(V) = 1$ with probability

$$r \triangleq \frac{(N+1)!}{(N+1)^{N+1}}. \tag{36}$$

Define

$$X_{i,H}(M,\mu) \triangleq \sum_{S(i)\cong H} \chi_\mu(V(S)) \prod_{(u,v)\in E(S)} M_{uv}. \tag{37}$$

Then $\mathbb{E}\left[X_{i,H}(M,\mu)\right] = rW_{i,H}(M)$, where $W_{i,H}(M)$ is defined in (1). Hence, $X_{i,H}(M,\mu)/r$ is an unbiased estimator of $W_{i,H}(M)$.

When H is a tree, the color coding together with the recursive tree structure enables us to use dynamic programming to count colorful trees and compute $X_{i,H}(M,\mu)$ efficiently. This is summarized as [32, Algorithm 2] for unrooted trees and the same algorithm with minor adjustments also works for rooted trees. First, since H is already a rooted tree, the step of assigning an arbitrary vertex of H as its root is not needed and thus the rooted

[8] In fact, the algorithm does not rely on the chandelier structure and works for any trees.

tree T_N constructed is exactly H itself. Second, as an intermediate step, [32, Algorithm 2] computes $Y(i, T_N, [N+1], \mu)$, which is the same as $\mathrm{aut}(H) X_{i,H}(M, \mu)$. Hence, we can simply output $\frac{1}{\mathrm{aut}(H)} Y(i, T_N, [N+1], \mu)$ as the rooted tree count $X_{i,H}(M, \mu)$.

Finally, we generate independent random colorings $\mu_1, \dots, \mu_t$ and average over $X_{i,H}(M, \mu_m)$'s to better approximate $W_{i,H}(M)$, where we set

$$t \triangleq \lceil 1/r \rceil .$$

Approximate similarity scores. To approximate $\Phi_{ij} \equiv \Phi_{ij}^{\mathcal{T}}$ in (3), we apply the above idea to each chandelier $H \in \mathcal{T}$. Generate $2t$ random colorings $\{\mu_a\}_{a=1}^t$ and $\{\nu_a\}_{a=1}^t$ which are independent copies of μ that map $[n]$ to $[N+1]$. Define

$$\widetilde{\Phi}_{ij} \triangleq \frac{1}{r^2} \sum_{H \in \mathcal{T}} \mathrm{aut}(H) \left(\frac{1}{t} \sum_{a=1}^t X_{i,H}(\overline{A}, \mu_a) \right) \left(\frac{1}{t} \sum_{a=1}^t X_{j,H}(\overline{B}, \nu_a) \right) . \tag{38}$$

Then $\mathbb{E}\left[\widetilde{\Phi}_{ij} \mid A, B\right] = \Phi_{ij}$. Moreover, the following result bounds the approximation error under the same conditions as those in Propositions 2 and 3 for the second moment calculation.

Proposition 4. *For any $i \in [n]$, if (20) holds,*

$$\frac{\mathrm{Var}[\widetilde{\Phi}_{ii} - \Phi_{ii}]}{\mathbb{E}[\Phi_{ii}]^2} = O\left(\frac{L^2}{\rho^2 nq} + \frac{L^2}{\rho^{2(K+M)} |\mathcal{J}|} \right) ; \tag{39}$$

for any $i \neq j$, if (22) holds,

$$\frac{\mathrm{Var}[\widetilde{\Phi}_{ij} - \Phi_{ij}]}{\mathbb{E}[\Phi_{ii}]^2} = O\left(\frac{1}{|\mathcal{T}| \rho^{2N}} \right) . \tag{40}$$

Finally, we show that the approximate similarity scores $\widetilde{\Phi}_{ij}$ can be computed efficiently using Algorithm 2.

Algorithm 2 Approximate similarity scores via color coding

1: **Input:** Centered adjacency matrices $\overline{A}$ and $\overline{B}$ and integers K, L, M, N, R.
2: Apply the algorithm for generating rooted trees in [7, Sec. 5] to list all non-isomorphic rooted trees with K edges, compute $\mathrm{aut}(H)$ for each rooted tree using the automorphism algorithm for trees in [11, Sec. 2], and return $\mathcal{J}$ as the subset of rooted trees whose number of automorphisms is at most R.
3: Generate (K, L, M, R)-chandeliers using $\mathcal{J}$ to obtain $\mathcal{T}$ per Definition 3.
4: Generate i.i.d. random colorings $\{\mu_a\}_{a=1}^t$ and $\{\nu_a\}_{a=1}^t$ mapping $[n]$ to $[N+1]$.
5: **for** each $a = 1, \cdots, t$ **do**
6: For each $H \in \mathcal{T}$, compute $\{X_{i,H}(\overline{A}, \mu_a)\}_{i \in [n]}$ and $\{X_{j,H}(\overline{B}, \nu_a)\}_{j \in [n]}$ via [32, Algorithm 2] with adjustments described after (37).
7: **end for**
8: **Output:** $\{\widetilde{\Phi}_{ij}\}_{i,j \in [n]}$ according to (38).

Proposition 5. *Algorithm 2 computes $\{\widetilde{\Phi}_{ij}\}_{i,j \in [n]}$ in time $O\left(n^2 (3e\alpha)^N\right)$. Furthermore, when $nq \geq 2$, under the choice of $K, L, M, R \in \mathbb{N}$ as per (10), the time complexity is $O(n^{c/\epsilon})$, where ϵ is from (10) and c is an absolute constant.*

Proof of Theorem 2. Note that

$$\begin{aligned} \mathrm{Var}[\widetilde{\Phi}_{ij}] &= \mathrm{Var}[\widetilde{\Phi}_{ij} - \Phi_{ij}] + \mathrm{Var}[\Phi_{ij}] + 2\mathrm{Cov}\left(\widetilde{\Phi}_{ij} - \Phi_{ij}, \Phi_{ij}\right) \\ &= \mathrm{Var}[\widetilde{\Phi}_{ij} - \Phi_{ij}] + \mathrm{Var}[\Phi_{ij}] , \end{aligned} \tag{41}$$

where the last equality holds because $\mathbb{E}\left[\widetilde{\Phi}_{ij} | A, B\right] = \Phi_{ij}$ and so

$$\mathrm{Cov}\left(\widetilde{\Phi}_{ij} - \Phi_{ij}, \Phi_{ij}\right) = \mathbb{E}\left[\mathbb{E}\left[(\widetilde{\Phi}_{ij} - \Phi_{ij}) | A, B\right] \Phi_{ij}\right] = 0 .$$

Under the assumption of Theorem 1, both (20) and (22) hold. Since $\mathbb{E}\left[\widetilde{\Phi}_{ij}\right] = \mathbb{E}\left[\Phi_{ij}\right]$, applying Proposition 4 yields

$$\frac{\mathrm{Var}[\widetilde{\Phi}_{ii}]}{\mathbb{E}\left[\widetilde{\Phi}_{ii}\right]^2} = O\left(\frac{L^2}{\rho^2 nq} + \frac{L^2}{\rho^{2(K+M)} |\mathcal{J}|} \right) ;$$

for all i and

$$\frac{\mathrm{Var}[\widetilde{\Phi}_{ij}]}{\mathbb{E}\left[\widetilde{\Phi}_{ii}\right]^2} = O\left(\frac{1}{|\mathcal{T}| \rho^{2N}} \right) .$$

for all $i \neq j$. In other words, Propositions 2– 3 and hence Theorem 1 continue to hold with $\widetilde{\Phi}_{ij}$ in place of Φ_{ij}. The time complexity follows from Proposition 5. □

5 SEEDED GRAPH MATCHING

Recall that with high probability Algorithm 1 applied to the class $\mathcal{T}$ of chandeliers finds a set I with $|I| = n - o(n)$ and recovers the latent permutation π on I. In this section, we develop a seeded graph matching subroutine (Algorithm 3) that matches the remaining vertices, thereby achieving exact recovery. Since the seed set I depends on graphs A and B, we need to show that Algorithm 3 succeeds even if the seed set I is chosen adversarially as long as $|I| = (1 - o(1))n$.

Given $I' \subset [n]$ and an injection $\pi' : I' \to [n]$, for any vertex i in A and vertex j in B, denote by $\mathsf{N}_{\pi'}(i, j)$ the number of common neighbors of i and j under the vertex correspondence π', namely, the number of vertex $u \in I'$ such that u is a neighbor of i in A and $\pi'(u)$ is a neighbor of j in B.

Algorithm 3 Seeded graph matching

1: **Input:** A and B, a mapping $\hat{\pi} : I \to [n]$, and γ.
2: Let $J = I$ and $\widetilde{\pi} = \hat{\pi}$.
3: **while** there exists $i \notin J$ and $j \notin \widetilde{\pi}(J)$ such that $\mathsf{N}_{\widetilde{\pi}}(i, j) \geq \gamma(n-2)q^2$ **do**
4: Add i to J and let $\widetilde{\pi}(i) = j$.
5: **end while**
6: **Output:** $\widetilde{\pi}$.

Algorithm 3 keeps adding vertices as new seeds once we are confident that they are true pairs based on the current seed set, in a similar fashion as the percolation graph matching proposed in [46]. It is a simplified version of [6, Algorithm 3.22], since our initial seeds are guaranteed to be error-free (thanks to Theorem 1) and so there is no need to clean up any mismatch. This allows us to show our Algorithm 3 succeeds under the information-theoretic necessary condition of $nq(q + \rho(1-q)) \geq (1+\epsilon) \log n$, whereas their algorithm requires $nq(q+\rho(1-q)) > \log^C n$ for some constant

$C > 1$. Another similar algorithm in prior work is [31, Algorithm 4], which however requires $nq \le \sqrt{n}/\log n$.

The following proposition gives sufficient conditions for our seeded algorithm to achieve exact recovery. Let

$$h(x) = x\log x - x + 1 \tag{42}$$

for $x > 0$, which is a convex function with the minimum value 0 achieved at $x = 1$.

Proposition 6. *Fix an arbitrarily small constant $\epsilon > 0$. Suppose $A, B \sim \mathcal{G}(n, q, \rho)$ with $q \le \frac{1}{2}$, $nq(q + \rho(1-q)) \ge (1+\epsilon)\log n$, and $\rho \ge \epsilon$. Let $\hat{\pi} \equiv \hat{\pi}(A, B)$ denote a mapping: $I \to [n]$ such that $\hat{\pi} = \pi|_I$ and $|I| \ge (1 - \epsilon/16)\, n$. Let γ denote the unique solution in $(1, +\infty)$ to $h(\gamma) = \frac{3\log n}{(n-2)q^2}$. Then with probability at least $1 - o(1)$, Algorithm 3 with inputs $\hat{\pi}$ and γ outputs $\widetilde{\pi} = \pi$ in $O(n^3q^2)$ time.*

Proof of Theorem 3. Theorem 1 ensures that, with probability $1 - o(1)$, Algorithm 1 returns a mapping $\hat{\pi} : I \to [n]$ in time $O(n^C)$ such that $\hat{\pi} = \pi|_I$ and $I \ge (1 - \epsilon/16)n$. Furthermore, Proposition 6 implies that, with probability $1 - o(1)$, Algorithm 3 outputs $\widetilde{\pi} = \pi$ in $O(n^3q^2)$ time. Hence, Theorem 3 follows. □

ACKNOWLEDGMENTS

C. Mao was supported in part by NSF Award DMS-2053333 and NSF Award DMS-2210734. Y. Wu was supported in part by NSF Award CCF-1900507, NSF CAREER Award CCF-1651588, and an Alfred Sloan fellowship. J. Xu was supported in part by NSF Award CCF-1856424 and NSF CAREER Award CCF-2144593. S. H. Yu was supported by NSF Award CCF-1856424. The authors are grateful for the hospitality and the support of the Simons Institute for the Theory of Computing at the University of California, Berkeley, where part of this work was carried out during the program on "Computational Complexity of Statistical Inference" in Fall 2021.

A AUXILIARY RESULTS

The following lemma computes the cross-moments of $\overline{A}_{uv}$ and $\overline{B}_{\pi(u)\pi(v)}$ from the centered adjacency matrices.

Lemma 1 ([32, Lemma 5]). *Let $(A, B) \sim \mathcal{G}(n, q, \rho)$. Assume $q \le \frac{1}{2}$. For any $0 \le \ell, m \le 2$ with $2 \le \ell + m \le 4$,*

$$\mathbb{E}\left[\sigma^{-\ell-m}\overline{A}_{uv}^{\ell}\overline{B}_{\pi(u)\pi(v)}^{m}\right] = \begin{cases} \rho^{\mathbf{1}\{\ell=m=1\}} & \ell+m=2 \\ \frac{\rho(1-2q)}{\sqrt{q(1-q)}} & \ell+m=3 \\ \frac{q(1-q)+\rho(1-2q)^2}{q(1-q)} & \ell+m=4 \end{cases}. \tag{43}$$

Moreover,

$$\left|\mathbb{E}\left[\sigma^{-\ell-m}\overline{A}_{uv}^{\ell}\overline{B}_{\pi(u)\pi(v)}^{m}\right]\right| \le |\rho|^{\mathbf{1}\{\ell=m=1\}}\mathbf{1}_{\{\ell+m=2\}} + \sqrt{\frac{1}{q}}\mathbf{1}_{\{\ell+m=3\}} + \frac{1}{q}\mathbf{1}_{\{\ell+m=4\}}. \tag{44}$$

B A DATA-DRIVEN CHOICE OF THE THRESHOLD

In this section, we describe a data-driven approach to choose threshold τ in Algorithm 1 without the knowledge of q and ρ. For each $i \in [n]$, let $\psi(i)$ denote one of the maximizer of Φ_{ij} over all $j \in [n]$. Let k denote the corresponding node such that $\Phi_{k\psi(k)}$ is the median of $\{\Phi_{i\psi(i)} : i \in [n]\}$. Set $\hat{\tau} = \frac{1}{2}\Phi_{k\psi(k)}$. We claim that $\frac{1}{2}c\mu \le \hat{\tau} \le \frac{1}{2}(2-c)\mu$ for any constant $0 < c < 1$ with probability $1 - o(1)$ when $nq = \omega(1)$ and with probability $1 - 3\delta$ for any constant $\delta \in (0, 1)$ when $nq \ge C(\epsilon, \delta)$. Hence by Theorem 1, $|I| = (1 - o(1))n$ with probability $1 - o(1)$ in the former case and $\mathbb{E}[|I|] \ge (1 - 3\delta)(1 - \delta)n \ge (1 - 4\delta)$ in the latter case.

It remains to show the claim, which reduces to proving $c\mu \le \Phi_{k\psi(k)} \le (2-c)\mu$. Without loss of generality, we assume $\pi = \mathrm{id}$. Let

$$J = \left\{ i \in [n] : i \in \arg\max_j \Phi_{ij} \text{ and } c\mu \le \Phi_{ii} \le (2-c)\mu \right\}.$$

Recall that $F = \{i : |\Phi_{ii} - \mu| > (1-c)\mu\}$ as defined in (31). By (34), with probability at least $1 - o(1)$, $\Phi_{ij} < c\mu$ for all $i \ne j$ and hence $J = [n]\backslash F$. Moreover, we have $\mathbb{E}[|F|] \le \gamma n$, where γ is given in (35). By Markov's inequality, $\mathbb{P}\{|F| \ge n/3\} \le 3\gamma$. Note that $\gamma = o(1)$ if $nq = \omega(1)$, and $\gamma < \delta$ for any constant $\delta \in (0, 1)$ if $nq \ge C(\epsilon, \delta)$. Hence, we have $|J| \ge 2n/3$ with probability $1 - o(1)$ if $nq = \omega(1)$, and with probability $1 - 3\delta$ if $nq \ge C(\epsilon, \delta)$. Henceforth assume $|J| \ge 2n/3$. If $\Phi_{k\psi(k)} > (2-c)\mu$, then there are at least $n/2$ nodes i with $\Phi_{i\psi(i)} > (2-c)\mu$, contradicting $|J| \ge 2n/3$. Analogous argument holds for the case of $\Phi_{k\psi(k)} < c\mu$. Thus, we must have $c\mu \le \Phi_{k\psi(k)} \le (2-c)\mu$.

REFERENCES

[1] Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. 2015. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences* 112, 10 (2015), 2942–2947. https://doi.org/10.1073/pnas.1401651112

[2] Noga Alon, Phuong Dao, Iman Hajirasouliha, Fereydoun Hormozdiari, and S Cenk Sahinalp. 2008. Biomolecular network motif counting and discovery by color coding. *Bioinformatics* 24, 13 (2008), i241–i249. https://doi.org/10.1093/bioinformatics/btn163

[3] Noga Alon, Raphael Yuster, and Uri Zwick. 1995. Color-coding. *Journal of the ACM (JACM)* 42, 4 (1995), 844–856. https://doi.org/10.1145/210332.210337

[4] Vikraman Arvind and Venkatesh Raman. 2002. Approximation algorithms for some parameterized counting problems. In *International Symposium on Algorithms and Computation*. Springer, 453–464. https://doi.org/10.1007/3-540-36136-7_40

[5] László Babai. 2016. Graph Isomorphism in Quasipolynomial Time [Extended Abstract]. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing* (Cambridge, MA, USA) *(STOC '16)*. ACM, New York, NY, USA, 684–697. https://doi.org/10.1145/2897518.2897542

[6] Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. 2019. (Nearly) Efficient Algorithms for the Graph Matching Problem on Correlated Random Graphs. In *Advances in Neural Information Processing Systems*. 9186–9194. https://doi.org/10.48550/arXiv.1805.02349

[7] Terry Beyer and Sandra Mitchell Hedetniemi. 1980. Constant time generation of rooted trees. *SIAM J. Comput.* 9, 4 (1980), 706–712. https://doi.org/10.1137/0209055

[8] Béla Bollobás. 1982. Distinguishing vertices of random graphs. *North-Holland Mathematics Studies* 62 (1982), 33–49. https://doi.org/10.1016/S0304-0208(08)73545-X

[9] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. 2016. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms* 49, 3 (2016), 503–532. https://doi.org/10.1002/rsa.20633

[10] Rainer E Burkard, Eranda Cela, Panos M Pardalos, and Leonidas S Pitsoulis. 1998. The quadratic assignment problem. In *Handbook of combinatorial optimization*. Springer, 1713–1809. https://doi.org/10.1007/978-1-4613-0303-9_27

[11] Charles J Colbourn and Kellogg S Booth. 1981. Linear time automorphism algorithms for trees, interval graphs, and planar graphs. *SIAM J. Comput.* 10, 1 (1981), 203–225. https://doi.org/10.1137/0210015

[12] Daniel Cullina and Negar Kiyavash. 2016. Improved achievability and converse bounds for Erdos-Rényi graph matching. *ACM SIGMETRICS performance evaluation review* 44, 1 (2016), 63–72. https://doi.org/10.1145/2964791.2901460

[13] Daniel Cullina and Negar Kiyavash. 2017. Exact alignment recovery for correlated Erdős-Rényi graphs. *arXiv 1711.06783* (2017). https://doi.org/10.48550/arXiv.1711.06783

[14] Daniel Cullina, Negar Kiyavash, Prateek Mittal, and H Vincent Poor. 2019. Partial Recovery of Erdős-Rényi Graph Alignment via k-Core Alignment. *Proceedings*

of the ACM on Measurement and Analysis of Computing Systems 3, 3 (2019), 1–21. https://doi.org/10.1145/3366702

[15] Tomek Czajka and Gopal Pandurangan. 2008. Improved random graph isomorphism. *Journal of Discrete Algorithms* 6, 1 (2008), 85–92. https://doi.org/10.1016/j.jda.2007.01.002

[16] Osman Emre Dai, Daniel Cullina, Negar Kiyavash, and Matthias Grossglauser. 2019. Analysis of a canonical labeling algorithm for the alignment of correlated Erdos-Rényi graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 2 (2019), 1–25. https://doi.org/10.1145/3341617.3326151

[17] Jian Ding and Hang Du. 2022. Matching recovery threshold for correlated random graphs. *arXiv preprint arXiv:2205.14650* (2022). https://doi.org/10.48550/arXiv.2205.14650

[18] Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. 2021. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields* 179, 1 (2021), 29–115. https://doi.org/10.1007/s00440-020-00997-4

[19] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. 2022. Spectral Graph Matching and Regularized Quadratic Relaxations I Algorithm and Gaussian Analysis. *Foundations of Computational Mathematics* (Jun 2022), 1–55. https://doi.org/10.1007/s10208-022-09570-y

[20] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. 2022. Spectral Graph Matching and Regularized Quadratic Relaxations II: Erdős-Rényi Graphs and Universality. *Foundations of Computational Mathematics* (Jun 2022), 1–51. https://doi.org/10.1007/s10208-022-09575-7

[21] Luca Ganassali and Laurent Massoulié. 2020. From tree matching to sparse graph alignment. In *Conference on Learning Theory*. PMLR, 1633–1665. https://doi.org/10.48550/arXiv.2002.01258

[22] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. 2021. Impossibility of partial recovery in the graph alignment problem. In *Conference on Learning Theory*. PMLR, 2080–2102. https://doi.org/10.48550/arXiv.2102.02685

[23] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. 2022. Correlation Detection in Trees for Planted Graph Alignment. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. 74:1–74:8. https://doi.org/10.4230/LIPIcs.ITCS.2022.74

[24] Luca Ganassali, Laurent Massoulié, and Guilhem Semerjian. 2022. Statistical limits of correlation detection in trees. *arXiv preprint arXiv:2209.13723* (2022). https://doi.org/10.48550/arXiv.2209.13723

[25] Georgina Hall and Laurent Massoulié. 2022. Partial recovery in the graph alignment problem. *Operations Research* (2022). https://doi.org/10.1287/opre.2022.2355

[26] Samuel B Hopkins and David Steurer. 2017. Efficient Bayesian Estimation from Few Samples: Community Detection and Related Problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, Los Alamitos, CA, USA, 379–390. https://doi.org/10.1109/FOCS.2017.42

[27] Camille Jordan. 1869. Sur les assemblages de lignes. *Journal für die reine und angewandte Mathematik* 70 (1869), 185–190. http://eudml.org/doc/148084

[28] Vince Lyzinski, Donniell Fishkind, Marcelo Fiori, Joshua Vogelstein, Carey Priebe, and Guillermo Sapiro. 2016. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 38, 1 (2016), 60–73. https://doi.org/10.1109/TPAMI.2015.2424894

[29] Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. 2010. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. In *International Colloquium on Automata, Languages, and Programming*. Springer, 594–604. https://doi.org/10.1007/978-3-642-14165-2_50

[30] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. 2021. Random Graph Matching with Improved Noise Robustness. In *Proceedings of Thirty Fourth Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 134)*. 3296–3329. https://doi.org/10.48550/arXiv.2101.11783

[31] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. 2023. Exact matching of random graphs with constant correlation. *Probability Theory and Related Fields* (2023), 1–63. https://doi.org/10.1007/s00440-022-01184-3

[32] Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. 2021. Testing network correlation efficiently via counting trees. *arXiv preprint arXiv:2110.11816* (2021). https://doi.org/10.48550/arXiv.2110.11816

[33] Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. 2022. Random graph matching at Otter's threshold via counting chandeliers. *arXiv preprint arXiv:2209.12313* (2022). https://doi.org/10.48550/arXiv.2209.12313

[34] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827. https://doi.org/10.1126/science.298.5594.824

[35] Elchanan Mossel, Joe Neeman, and Allan Sly. 2015. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields* 162, 3 (2015), 431–461. https://doi.org/10.1007/s00440-014-0576-6

[36] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125. https://doi.org/10.1109/SP.2008.33

[37] Christoffer Olsson and Stephan Wagner. 2022. Automorphisms of random trees. In *33rd International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.AofA.2022.16

[38] Richard Otter. 1948. The number of trees. *Annals of Mathematics* (1948), 583–599. https://doi.org/10.2307/1969046

[39] Pedram Pedarsani and Matthias Grossglauser. 2011. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1235–1243. https://doi.org/10.1145/2020408.2020596

[40] Giovanni Piccioli, Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. 2022. Aligning random graphs with a sub-tree similarity message-passing algorithm. *Journal of Statistical Mechanics: Theory and Experiment* 2022, 6 (2022), 063401. https://doi.org/10.1088/1742-5468/ac70d2

[41] George Pólya. 1937. Kombinatorische anzahlbestimmungen für gruppen, graphen und chemische verbindungen. *Acta mathematica* 68 (1937), 145–254.

[42] Pedro Ribeiro, Pedro Paredes, Miguel EP Silva, David Aparicio, and Fernando Silva. 2021. A survey on subgraph counting: concepts, algorithms, and applications to network motifs and graphlets. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36. https://doi.org/10.1145/3433652

[43] Rohit Singh, Jinbo Xu, and Bonnie Berger. 2008. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* 105, 35 (2008), 12763–12768. https://doi.org/10.1073/pnas.0806627105

[44] Shinji Umeyama. 1988. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10, 5 (1988), 695–703. https://doi.org/10.1109/34.6778

[45] Yihong Wu, Jiaming Xu, and Sophie H. Yu. 2022. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory* 68, 8 (Apr 2022), 5391–5417. https://doi.org/10.1109/TIT.2022.3169005

[46] Lyudmila Yartseva and Matthias Grossglauser. 2013. On the performance of percolation graph matching. In *Proceedings of the first ACM conference on Online social networks*. 119–130. https://doi.org/10.1145/2512938.2512952

[47] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. 2008. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 12 (2008), 2227–2242. https://doi.org/10.1109/TPAMI.2008.245

Received 2022-11-07; accepted 2023-02-06