Probability Theory and Related Fields
https://doi.org/10.1007/s00440-022-01184-3

Exact matching of random graphs with constant correlation

Cheng Mao[1] · **Mark Rudelson**[2] · **Konstantin Tikhomirov**[1]

Received: 20 October 2021 / Revised: 6 July 2022 / Accepted: 17 December 2022

Abstract
This paper deals with the problem of *graph matching* or *network alignment* for Erdős–Rényi graphs, which can be viewed as a noisy average-case version of the graph isomorphism problem. Let G and G' be $G(n, p)$ Erdős–Rényi graphs marginally, identified with their adjacency matrices. Assume that G and G' are correlated such that $\mathbb{E}[G_{ij}G'_{ij}] = p(1-\alpha)$. For a permutation π representing a latent matching between the vertices of G and G', denote by G^π the graph obtained from permuting the vertices of G by π. Observing G^π and G', we aim to recover the matching π. In this work, we show that for every $\varepsilon \in (0, 1]$, there is $n_0 > 0$ depending on ε and absolute constants $\alpha_0, R > 0$ with the following property. Let $n \geq n_0$, $(1+\varepsilon)\log n \leq np \leq n^{\frac{1}{R\log\log n}}$, and $0 < \alpha < \min(\alpha_0, \varepsilon/4)$. There is a polynomial-time algorithm F such that $\mathbb{P}\{F(G^\pi, G') = \pi\} = 1 - o(1)$. This is the first polynomial-time algorithm that recovers the *exact matching* between vertices of correlated Erdős–Rényi graphs with *constant correlation* with high probability. The algorithm is based on comparison of *partition trees* associated with the graph vertices.

Mathematics Subject Classification 05C85 · 05C80

C.M. was partially supported by NSF grant DMS-2053333. M.R. was partially supported by the NSF grant DMS-2054408. K.T. was partially supported by the Sloan Research Fellowship and by NSF grant DMS-2054666.

✉ Konstantin Tikhomirov
ktikhomirov6@gatech.edu

Cheng Mao
cheng.mao@math.gatech.edu

Mark Rudelson
rudelson@umich.edu

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, USA

[2] Department of Mathematics, University of Michigan, Ann Arbor, USA

Published online: 06 January 2023

Contents

1 Introduction

The problem of *graph matching* (also known as *graph alignment* or *network alignment*) refers to finding a mapping between vertices of two given graphs in order to maximize alignment of their edges. If the two graphs are isomorphic, the problem is the celebrated *graph isomorphism problem*, for which no polynomial-time algorithm is known in the worst case (see [1] and references therein). In general, the noisy graph matching problem can be formulated as the *quadratic assignment problem*, which is NP-hard to solve or approximate (see surveys [3, 23]).

While in the worst case the problem appears intractable, an optimal matching of certain *random graphs* can be realized in polynomial time. In particular, the graph isomorphism problem for Erdős–Rényi graphs above the connectivity threshold can be solved in polynomial time with high probability [4, 5, 9]. More recently, numerous results have been obtained in the literature for matching a pair of *correlated Erdős–Rényi graphs* [2, 6–8, 10, 11, 13–15, 17, 18, 20–22, 24, 26, 27]. At the same time, conditions for existence of a polynomial-time algorithm for recovering the latent matching between the two graphs are far from being fully understood. In this work, we make further progress along this line of research by proposing a polynomial-time algorithm which produces an exact matching between a pair of correlated Erdős–Rényi graphs with *constant correlation*, which is the first result of this kind in the literature.

1.1 The correlated Erdos–Rényi graph model

We consider the *correlated Erdős–Rényi graph model* [22] in this work. Fix $p \in (0, 1)$, $\alpha \in [0, 1 - p]$, and a positive integer n. Let G_0 be a $G(n, \frac{p}{1-\alpha})$ Erdős–Rényi graph, which is called the *parent graph*. Conditional on the parent graph G_0, a subgraph G is obtained by removing every edge of G_0 independently with probability α; moreover, another subgraph G' of G_0 is obtained in the same way (conditionally) independently of G. Then G and G' are marginally both $G(n, p)$ graphs, and for every pair of distinct vertices i and j in $[n] := \{1, 2, \ldots, n\}$,

$$\mathbb{P}\{i \text{ is adjacent to } j \text{ in } G \mid i \text{ is adjacent to } j \text{ in } G'\} = 1 - \alpha.$$

Note that α indicates the noise level in the model, while $1 - \alpha$ can be viewed as the correlation between the two graphs. Given a permutation $\pi : [n] \to [n]$, let G^π denote the graph obtained from permuting the vertices of G by π. In other words, i is adjacent to j in G if and only if $\pi(i)$ is adjacent to $\pi(j)$ in G^π. The permutation π is unknown and represents the latent matching between the vertices of the two graphs. Observing the graphs G^π and G', we aim to recover the matching π exactly.

1.2 Prior work and our contributions

We use the standard asymptotic notation $O(\cdot)$, $o(\cdot)$, and $\Omega(\cdot)$ for a growing n; we also use $\tilde{O}(\cdot)$ to hide a polylogarithmic factor in n. Moreover, we use C, C', c, c', possibly with subscripts, to denote universal positive constants that may change at each appearance.

Let us focus our discussion on the exact recovery of the latent matching π. First, it is without loss of generality to assume that the average degree of each graph exceeds the so-called *connectivity threshold*. To be more precise, if $np \le (1 - \varepsilon) \log n$, a $G(n, p)$ graph will almost surely contain isolated vertices, so exact recovery of the matching is impossible in this case. We therefore assume that the average degree satisfies $np \ge (1 + \varepsilon) \log n$ for an arbitrarily small absolute constant $\varepsilon > 0$. Then a $G(n, p)$ graph is known to be connected almost surely as n grows.

For the correlated Erdős–Rényi graph model, the optimal information-theoretic threshold for exact recovery of π is known [26]. For example, in the regime $\frac{p}{1-\alpha} = o(1)$, exact matching is possible if $np(1-\alpha) \ge (1+\varepsilon') \log n$ for any constant $\varepsilon' > 0$. In particular, if $np = (1 + \varepsilon) \log n$ for a small constant $\varepsilon > 0$, then this threshold requires α to be slightly smaller than ε. In the dense case where np is much larger than $\log n$, the information-theoretic threshold even allows α to be close to 1. However, this optimal condition is achieved by the maximum likelihood estimator which employs an exhaustive search over the set of permutations and is therefore computationally infeasible. Several recent works developed quasi-polynomial and polynomial time algorithms for exact recovery of π under stronger conditions. A selection of prior results along with ours are listed in Table 1.

As shown in Table 1, before this work, no polynomial-time algorithm is known to achieve exact recovery if the noise parameter α is a small constant nor if the average

Table 1 Conditions for exact matching

	Condition	Time complexity
[26]	$np(1-\alpha) \geq (1+\varepsilon')\log n$ if $\frac{p}{1-\alpha} = o(1)$	Exponential
[2]	$np \geq n^{o(1)}$, $1-\alpha \geq (\log n)^{-o(1)}$	$n^{O(\log n)}$
[11]	$np \geq (\log n)^C$, $\alpha \leq (\log n)^{-C}$	$\tilde{O}(n^3p^2 + n^{2.5})$
	$C\log n \leq np \leq e^{(\log\log n)^C}$, $\alpha \leq (\log\log n)^{-C}$	
[12]	$np \geq (\log n)^C$, $\alpha \leq (\log n)^{-C}$	$O(n^3)$
[20]	$np \geq (\log n)^C$, $\alpha \leq (\log\log n)^{-C}$	$\tilde{O}(n^2)$
This work	$(1+\varepsilon)\log n \leq np \leq n^{\frac{1}{C\log\log n}}$, $\alpha \leq \min(\text{const}, \varepsilon/4)$	$n^{2+o(1)}$

degree np is close to the connectivity threshold $\log n$. Our work achieves both conditions and therefore resolves what was seen as a main open problem in this literature. In particular, if $np = (1+\varepsilon)\log n$ for a constant $\varepsilon > 0$, the condition required by our algorithm differs from the optimal information-theoretic condition by at most a constant factor.

While our main focus is exact recovery of the latent matching π, part of our strategy applies to partial recovery of π and is expected to carry over to sparser regimes where exact matching is impossible; see Theorem A and Sect. 2.2 for details. The new algorithm we propose is based on exploring large neighborhoods of vertices via *partition trees*, a technique that may be of further interest. Moreover, the last step of our algorithm is to obtain an exact matching from a (potentially adversarial) partial matching. To this end, we develop a method that tolerates any constant fraction of wrongly matched pairs in the initial partial matching; see Sects. 2.3 and 7 for details.

1.3 Main results

Everywhere in this paper, when discussing the computational complexity of a function, we assume that elementary arithmetic operations as well as the square root and the logarithm, can be computed exactly in time $\tilde{O}(1)$. We note that analyzing the algorithms using the floating point arithmetic, while certainly possible, adds unnecessary technical details to the presentation and does not affect the order of the time complexity.

Our first main result deals with almost exact recovery of the latent permutation. The algorithm succeeds with probability at least $1 - n^{-D}$ for an arbitrary constant $D > 0$, even when the average degree is logarithmic in n.

Theorem A (Almost exact matching). *For any constant $D > 0$, there exist constants $\alpha_0, n_0, R, c > 0$ depending on D with the following property. Let G^π and G' be the graphs given by the correlated Erdős–Rényi graph model defined in Sect. 1.1, with parameters n, p, and α such that*

$$n \geq n_0, \qquad \alpha \in (0, \alpha_0), \qquad \log n \leq np(1-\alpha) \leq n^{\frac{1}{R\log\log n}}.$$

Then there is a random function F_{al} defined on pairs of graphs on $[n]$ and taking values in the set of permutations on $[n]$, such that

- *F_{al} is independent from the graphs G^π and G',*
- *F_{al} has expected time complexity $\tilde{O}(n^2)$, and*
- *for any latent permutation $\pi : [n] \to [n]$,*

$$\mathbb{P}\big\{F_{al}(G^\pi, G')(i) \neq \pi(i) \text{ for at most } n^{1-c} \text{ indices } i \in n\big\} \geq 1 - n^{-D}.$$

The next result establishes existence of a polytime procedure producing an exact matching with high probability.

Theorem B (Exact matching). *For any constant $\varepsilon \in (0, 1]$, there exists a constant $n_0 > 0$ depending on ε and absolute constants α_0, $R > 0$ with the following property. Let G^π and G' be the graphs given by the correlated Erdős–Rényi graph model defined in Sect. 1.1, with parameters n, p, and α such that*

$$n \geq n_0, \qquad (1+\varepsilon)\log n \leq np \leq n^{\frac{1}{R\log\log n}}, \qquad 0 < \alpha \leq \min(\alpha_0, \varepsilon/4).$$

Then there is a random function F_{ex} defined on pairs of graphs on $[n]$ and taking values in the set of permutations on $[n]$, such that

- *F_{ex} is independent from the graphs G^π and G',*
- *F_{ex} has expected time complexity $n^{2+o(1)}$, and*
- *for every permutation $\pi : [n] \to [n]$,*

$$\mathbb{P}\big\{F_{ex}(G^\pi, G') = \pi\big\} \geq 1 - n^{-10} - \exp(-\varepsilon pn/10).$$

The actual computational procedures for F_{al} and F_{ex} in the above theorems will be discussed in Sect. 2. Note that the theorems assume $np \leq n^{\frac{1}{R\log\log n}}$, that is, the graphs in consideration are sufficiently sparse. This is because the success of our main algorithm relies on the condition that the neighborhood of radius $O(\log\log n)$ around any typical vertex is a tree. In the denser regime where $n^{\frac{1}{R\log\log n}} \ll np \leq O(1)$, the problem of matching two Erdős–Rényi graphs with constant correlation remains open. It is interesting to study whether an extension of the algorithms in this work or our earlier work [20] can solve the problem. A major difficulty is to handle the probabilistic dependency across multiple steps of an iterative algorithm in the dense regime.

1.4 Notation

For any positive integer n, let $[n]$ be the set of integers $\{1, 2, \ldots, n\}$. Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{N}_0$ the set of nonnegative integers. Let $\wedge$ and $\vee$ denote the min and the max operator for two real numbers, respectively.

For a graph G with vertex set $[n]$ and $i \in [n]$, let $\deg_G(i)$ denote the degree of i in G. For distinct vertices $i, j \in [n]$, let $\mathrm{dist}_G(i, j)$ denote the distance between i and j

in the graph G. Let $\mathcal{N}_G(i)$ denote the set of neighbors of i in G. For a subset $S \subset [n]$, let $\mathcal{N}_G(S) := \bigcup_{i\in S} \mathcal{N}_G(i)$. For $r \in \mathbb{N}$, let $\mathcal{B}_G(i, r)$ and $\mathcal{S}_G(i, r)$ denote the ball and the sphere of radius r centered at i in G, respectively. Let $G(S)$ denote the subgraph of G induced by $S \subset [n]$. In particular, $G(\mathcal{B}_G(i, r))$ is the *r-neighborhood* of i in G.

2 Algorithms and theoretical guarantees

Let $n \in \mathbb{N}$ and $p \in (0, 1)$ be global constants that are known to the algorithms.

2.1 Partition trees and vertex signatures

In order to recover the latent matching π between vertices of the two graphs, we associate a *signature*, that is, a 2^m–dimensional vector, to every vertex in G^π and every vertex in G'. The signature of vertex i in a given graph Γ is constructed based on the *partition tree* rooted at i, which is, by definition, a complete binary tree T whose nodes $\{T_s^m\}_{s\in\{-1,1\}^m}$ at level m form a partition of the sphere $\mathcal{S}_\Gamma(i, m)$. Algorithm 1 gives the precise construction of the partition tree and the signature associated to a vertex.

Algorithm 1 `VertexSignature`

Input: a graph Γ on the vertex set $[n]$, a vertex $i \in [n]$, and a depth parameter $m \in \mathbb{N}$
Output: a signature vector $f \in \mathbb{R}^{2^m}$ and a vector of variances $\mathsf{v} \in \mathbb{R}^{2^m}$

$T_\varnothing^0 \leftarrow \{i\}$ ▷ $\varnothing$ *denotes the empty tuple*
for $k = 0, \ldots, m-1$ **do**
  for $s \in \{-1, 1\}^k$ **do**
    $T_{(s,+1)}^{k+1} \leftarrow \left\{j \in \mathcal{N}_\Gamma(T_s^k) \cap \mathcal{S}_\Gamma(i, k+1) : \deg_\Gamma(j) \geq np\right\}$
    $T_{(s,-1)}^{k+1} \leftarrow \left\{j \in \mathcal{N}_\Gamma(T_s^k) \cap \mathcal{S}_\Gamma(i, k+1) : \deg_\Gamma(j) < np\right\}$
  end for
end for
define $f(i) \in \mathbb{R}^{2^m}$ by $f(i)_s := \sum_{j\in\mathcal{N}_\Gamma(T_s^m)\cap\mathcal{S}_\Gamma(i,m+1)} \left(\deg_\Gamma(j) - 1 - np\right)$ for $s \in \{-1, 1\}^m$
define $\mathsf{v}(i) \in \mathbb{R}^{2^m}$ by $\mathsf{v}(i)_s := np(1-p)|\mathcal{N}_\Gamma(T_s^m)\cap\mathcal{S}_\Gamma(i, m+1)|$ for $s \in \{-1, 1\}^m$
return $f(i)$ and $\mathsf{v}(i)$

Algorithm 1 can be informally described as follows. Given a vertex i, we construct inductively a binary tree of sets T_s^k, $k = 0, \ldots, m$, $s \in \{-1, 1\}^k$, starting with $T_\varnothing^0 := \{i\}$. Each set T_s^k, $s \in \{-1, 1\}^k$, is a subset of the sphere $\mathcal{S}_\Gamma(i, k)$. For every $k < m$ and $s \in \{-1, 1\}^k$, T_s^k has two children $T_{(s,-1)}^{k+1}$ and $T_{(s,+1)}^{k+1}$, with the union $T_{(s,-1)}^{k+1} \cup T_{(s,+1)}^{k+1}$ equal to $\mathcal{N}_\Gamma(T_s^k)\cap\mathcal{S}_\Gamma(i, k+1)$. Here, we write $(s, \pm 1)$ for a binary vector in $\{-1, 1\}^{k+1}$ formed by concatenating s and ± 1. The set $T_{(s,-1)}^{k+1}$ is the collection of all vertices in $\mathcal{N}_\Gamma(T_s^k)\cap\mathcal{S}_\Gamma(i, k+1)$ with degree strictly less than np, and $T_{(s,+1)}^{k+1}$ — the vertices with degrees at least np (note that the input of the algorithm is a realization of a $G(n, p)$

$$T^0_{\varnothing} = \{i\};\ \ T^1_{(-1)} = \{i_1, i_2\};\ \ T^1_{(+1)} = \{i_3\};$$

$$T^2_{(-1,-1)} = \{i_{12}, i_{21}, i_{22}\};\ \ T^2_{(-1,+1)} = \{i_{11}\};\ \ T^2_{(+1,-1)} = \{i_{31}, i_{32}\};\ \ T^2_{(+1,+1)} = \{i_{33}\}.$$

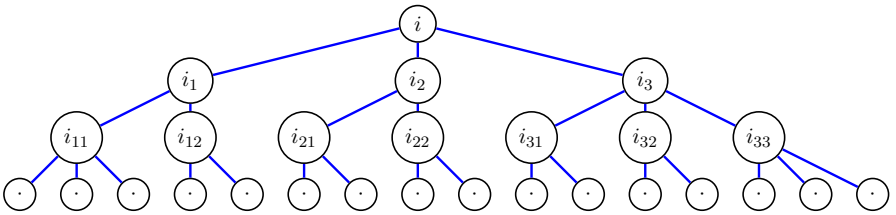

Fig. 1 An example of a partition tree of a vertex i of a graph with parameters $np = 3.5$ and $m = 2$. The blue lines denote the edges of the graph. The nodes of the partition tree of i are $T^0_{\varnothing} = \{i\};\ \ T^1_{(-1)} = \{i_1, i_2\};\ \ T^1_{(+1)} = \{i_3\}; T^2_{(-1,-1)} = \{i_{12}, i_{21}, i_{22}\};\ \ T^2_{(-1,+1)} = \{i_{11}\};\ \ T^2_{(+1,-1)} = \{i_{31}, i_{32}\};\ \ T^2_{(+1,+1)} = \{i_{33}\}$ (color figure online)

random graph, hence the threshold value). The collection of sets of vertices

$$T = \left\{T^k_s : k = 0, \ldots, m,\ s \in \{-1, 1\}^k\right\}$$

associated with a given vertex i is referred to as the partition tree rooted at i. It can be viewed as a data structure encoding statistics of *paths* of length m starting at i and classified according to the degrees of the comprised vertices. The key point of our approach is that the partition trees contain sufficient information for recovering the latent matching between the correlated graphs. We refer to Fig. 1 for an example of the partition tree of a vertex in a graph.

With the partition tree constructed, Algorithm 1 then defines the signature $f(i) \in \mathbb{R}^{2^m}$ of vertex i to be a vector whose entries are based on degrees of neighbors of vertices in the leaves of the partition tree. Finally, the auxiliary vector $\mathsf{v}(i) \in \mathbb{R}^{2^m}$ encodes the variances of the entries of the signature vector in a $G(n, p)$ random graph, conditional on a realization of the $(m+1)$–neighborhood of i. For matching vertices of G^{π} and G', we will use normalized differences of signatures with components $\frac{f_s(i) - f'_s(i')}{\sqrt{\mathsf{v}_s(i) + \mathsf{v}'_s(i')}}$, where the superscript "$'$" denotes that the signature vector and the vector of variances are for a vertex i' in G'; see Algorithm 2 for details.

Since we will take $m = O(\log\log n)$ and the average degree of each graph is assumed to be $n^{\frac{1}{R\log\log n}}$ for a sufficiently large constant R, the expected time complexity of computing one signature vector with Algorithm 1 given the adjacency matrix of Γ is $O(n)$.

2.2 Almost exact matching

With the signatures constructed, we then match vertex i in G^{π} and vertex i' in G' if and only if their signatures are sufficiently close. Ideally, the difference between the signatures of a "correct" pair of vertices, $\pi(i)$ in G^{π} and i in G', should be small,

while the difference between the signatures of a "wrong" pair of vertices, $\pi(i)$ in G^π and i' in G' for $i \neq i'$, should be large. Algorithm 2 compares signatures in terms of a *sparsified* ℓ_2–distance weighted by the associated variances. The results of the vertex comparisons are stored in an $n \times n$ matrix, denoted by B in the algorithm description.

Algorithm 2 `SignatureComparison`

Input: two graphs Γ and Γ' on the vertex set $[n]$
Output: a matrix $B \in \{0,1\}^{n\times n}$

$m \leftarrow \lceil 22 \log\log n \rceil$
$w \leftarrow \lfloor (\log n)^5 \rfloor$
for $i = 1, \ldots, n$ **do**
  $\big(f(i), \mathsf{v}(i)\big) \leftarrow$ `VertexSignature`(Γ, i, m)
  $\big(f'(i), \mathsf{v}'(i)\big) \leftarrow$ `VertexSignature`(Γ', i, m)
end for
$J \leftarrow$ a uniform random subset of $\{-1,1\}^m$ of cardinality $2w$
for $i = 1, \ldots, n$ **do**
  for $i' = 1, \ldots, n$ **do**
    if $\sum_{s\in J} \frac{(f_s(i)-f'_s(i'))^2}{\mathsf{v}_s(i)+\mathsf{v}'_s(i')} < 2w\big(1 - \frac{1}{\sqrt{\log n}}\big)$ **then**
      $B_{i,i'} \leftarrow 1$
    else
      $B_{i,i'} \leftarrow 0$
    end if
  end for
end for
return B

To be more precise, given two vertices i and i' in graphs Γ and Γ', with signatures $f(i)$ and $f'(i')$ and variance vectors $\mathsf{v}(i)$ and $\mathsf{v}'(i')$, respectively, the algorithm computes the sum $\sum_{s\in J} \frac{(f_s(i)-f'_s(i'))^2}{\mathsf{v}_s(i)+\mathsf{v}'_s(i')}$, where J is a uniform random subset of $\{-1,1\}^m$ of a cardinality polylogarithmic in n. If this sum is smaller than the threshold $|J|\big(1-\frac{1}{\sqrt{\log n}}\big)$, then we match the vertices i and i'. The main difficulty of the signature comparison is that, under the assumption of constant correlation between the graphs G and G', the signature vectors of vertex i in G and G' will be only slightly correlated with a high probability. To distinguish between a correct and a wrong matching, we need to be able to distinguish between "very slightly correlated" and "essentially uncorrelated" signature vectors, which is achieved through a rather delicate analysis. This is the reason why the threshold $|J|\big(1 - \frac{1}{\sqrt{\log n}}\big)$ is only slightly different from the value $|J|$ which would be the expected squared ℓ_2–distance between two independent random vectors in R^J, normalized so that the variance of each component of the difference is one.

Moreover, taking the sparsified distance over J in Algorithm 2, rather than summing over all indices $s \in \{-1,1\}^m$, weakens the dependence across entries of the signature vectors and allows us to prove strong concentration bounds for $\sum_{s\in J} \frac{(f_s(i)-f'_s(i'))^2}{\mathsf{v}_s(i)+\mathsf{v}'_s(i')}$.

The idea of comparing the sparsified signature vectors is taken from earlier work [20] by the current authors.

It is not difficult to see that the expected time complexity of Algorithm 2 is of order $\tilde{O}(n^2)$, because it amounts to computing and comparing all the signature vectors, which are of length polylogarithmic in n. Theorem 2.1, which is the main technical result of the paper, then guarantees that Algorithm 2 distinguishes correct pairs from wrong pairs for most vertices with high probability. The proof of the theorem is given at the end of Sect. 6.

Theorem 2.1 (Difference between signatures of typical vertices). *For any constant $D > 0$, there exist constants $\alpha_0, n_0, R, c > 0$ depending on D with the following property. Let G^π and G' be the two graphs given by the correlated Erdős–Rényi graph model defined in Sect. 1.1 with underlying matching $\pi : [n] \to [n]$ and parameters n, p, and α such that*

$$n \geq n_0, \qquad \alpha \in (0, \alpha_0), \qquad \log n \leq np(1-\alpha) \leq n^{\frac{1}{R \log\log n}}.$$

Let $B \in \{0,1\}^{n \times n}$ be given by Algorithm 2 with G^π and G' as input graphs. Then, with probability at least $1 - n^{-D}$, there exists a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that $B_{\pi(i),i} = 1$ for any $i \in \mathcal{I}$, and $B_{\pi(i),i'} = 0$ for any distinct $i, i' \in \mathcal{I}$.

To pass from a matrix B in the above theorem to a permutation, we apply a simple procedure, Algorithm 3, which yields an almost exact estimate $\hat{\pi}$ of the underlying matching π. The computational complexity of Algorithm 3 is clearly $O(n^2)$. Proposition 2.2 ensures that Algorithm 3 succeeds deterministically.

Algorithm 3 `ApproximateMatching`

Input: a binary matrix $B \in \{0,1\}^{n \times n}$
Output: a permutation $\hat{\pi} : [n] \to [n]$

$H \leftarrow$ the bipartite graph whose adjacency matrix is B
let $V = V' = [n]$ be the two parts of vertices of H
while the edge set of H is nonempty **do**
  pick an arbitrary edge $i \sim i'$ in H where $i \in V$ and $i' \in V'$
  define $\hat{\pi}(i') := i$
  delete the edge $i \sim i'$ from H
  $V \leftarrow V \backslash \{i\}$
  $V' \leftarrow V' \backslash \{i'\}$
end while
if $V \neq \varnothing$ **then**
  define $\hat{\pi}|_{V'}$ to be an arbitrary bijection from V' to V so that $\hat{\pi}$ is a permutation on $[n]$
end if
return $\hat{\pi}$

Proposition 2.2 (Matching from comparisons). *Fix a permutation $\pi : [n] \to [n]$, a matrix $B \in \{0,1\}^{n\times n}$, and a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - k$ for a positive integer $k \leq n/4$. Suppose that $B_{\pi(i),i} = 1$ for any $i \in \mathcal{I}$, and $B_{\pi(i),i'} = 0$ for any distinct $i, i' \in \mathcal{I}$. Then Algorithm 3 outputs a permutation $\hat{\pi} : [n] \to [n]$ satisfying*

$$\left|\left\{i \in [n] : \hat{\pi}(i) \neq \pi(i)\right\}\right| \leq 4k.$$

Proof First, it is clear that Algorithm 3 outputs a bijection $\hat{\pi} : [n] \to [n]$, so $\hat{\pi}$ is well-defined.

Next, we claim that the while loop in Algorithm 3 will be run for at least $n - 2k$ iterations. To this end, suppose that it has been run for strictly fewer than $n - 2k$ iterations, after which we have $|V| = |V'| \geq 2k + 1$. Since $|\mathcal{I}| \geq n - k$, it follows that $|\mathcal{I} \cap V'| \geq k + 1$. For each $i' \in \mathcal{I} \cap V'$, consider two cases:

- Suppose that $\pi(i') \in V$, that is, $\pi(i')$ has not been deleted. By the assumption on B, we have $B_{\pi(i'),i} = 1$, so the edge $\pi(i') \sim i'$ is still present in the graph H. As a result, the while loop will be run for at least one more iteration.
- Suppose that $\pi(i')$ has already been deleted in a previous iteration, say, along with another vertex j'. Then there is an edge $\pi(i') \sim j'$ in the original bipartite graph, that is, $B_{\pi(i'),j'} = 1$. As $j' \neq i'$, by the assumption on B, we must have $j' \in [n]\backslash\mathcal{I}$. Since $\left|[n]\backslash\mathcal{I}\right| \leq k$, this case can occur for at most k vertices i'.

To conclude, because $|\mathcal{I} \cap V'| \geq k + 1$, there is at least one $i' \in \mathcal{I} \cap V'$ that falls into the first case above. Thus, the while loop will be run for at least one more iteration, and the claim is proved.

Furthermore, consider an iteration of the while loop in which we pick an edge $i \sim i'$ in H for $i \in V$ and $i' \in V'$. There are two cases:

- Suppose that both $\pi^{-1}(i)$ and i' are in $\mathcal{I}$. Since $B_{i,i'} = 1$, by the assumption on B, we must have $\pi^{-1}(i) = i'$ so that $\pi(i') = i = \hat{\pi}(i')$.
- Suppose that either $\pi^{-1}(i)$ or i' is in $[n]\backslash\mathcal{I}$. After this iteration, i and i' are deleted from the vertex sets. Since $\left|[n]\backslash\mathcal{I}\right| \leq k$, this case can occur at most $2k$ times in total.

Recall that the while loop will be run for at least $n - 2k$ iterations, and among them, at least $n - 4k$ iterations fall into the first case above. Consequently, we have $\pi(i') = \hat{\pi}(i')$ for at least $n - 4k$ vertices $i' \in [n]$. □

Combining Theorem 2.1 and Proposition 2.2 immediately yields the following result.

Corollary 2.3 (Almost exact matching). *In the same setting as in Theorem 2.1, using the matrix $B \in \{-1,1\}^{n\times n}$ given by Algorithm 2 as the input, we run Algorithm 3 to produce $\hat{\pi} : [n] \to [n]$. Then, with probability at least $1 - n^{-D}$, it holds that*

$$\left|\left\{i \in [n] : \hat{\pi}(i) \neq \pi(i)\right\}\right| \leq 4n^{1-c}.$$

Note that Corollary 2.3 implies Theorem A in the introduction.

2.3 Exact matching

Having obtained an estimate $\hat{\pi}$ of the underlying matching π, we now aim to recover π exactly by refining $\hat{\pi}$. The algorithm we propose is based on iterative refinements of an initial partial matching by studying intersections of neighborhoods of vertices in G^{π} and G'. At each step, we obtain a matching with the number of incorrectly matched pairs of vertices smaller by a constant factor than that number in the previous step. After a logarithmic number of such iterations, we obtain the exact matching with high probability. A formal description of the procedure is given in Algorithm 4.

Algorithm 4 `RefinedMatching`

Input: two graphs Γ and Γ' on $[n]$, a permutation $\hat{\pi} : [n] \to [n]$, and a parameter $\varepsilon > 0$

Output: a permutation $\tilde{\pi} : [n] \to [n]$

$\pi_0 \leftarrow \hat{\pi}$
for $\ell = 1, \ldots, \lceil \log_2 n \rceil$ **do**
  for $i = 1, \ldots, n$ **do**
    if there is a vertex $i' \in [n]$ such that
      • $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{\Gamma}(i)\right) \cap \mathcal{N}_{\Gamma'}(i')\right| \geq \varepsilon^2 pn/512$
      • $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{\Gamma}(i)\right) \cap \mathcal{N}_{\Gamma'}(j')\right| < \varepsilon^2 pn/512$ for all $j' \in [n]\backslash\{i'\}$
      • $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{\Gamma}(j)\right) \cap \mathcal{N}_{\Gamma'}(i')\right| < \varepsilon^2 pn/512$ for all $j \in [n]\backslash\{i\}$
    then
      $\pi_{\ell}(i') \leftarrow i$
    end if
  end for
  extend π_{ℓ} to a permutation on $[n]$ in an arbitrary way
end for
$\tilde{\pi} \leftarrow \pi_{\lceil \log_2 n \rceil}$
return $\tilde{\pi}$

The underlying reason for why Algorithm 4 succeeds is a certain expansion property of sparse Erdős–Rényi graphs. Note that at each step of Algorithm 4, we assign $\pi_{\ell}(i') := i$ whenever i' is a vertex in $[n]$ with $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{G^{\pi}}(i)\right) \cap \mathcal{N}_{G'}(i')\right|$ "large" and both $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{G^{\pi}}(i)\right) \cap \mathcal{N}_{G'}(j')\right|$ and $\left|\pi_{\ell-1}^{-1}\left(\mathcal{N}_{G^{\pi}}(j)\right) \cap \mathcal{N}_{G'}(i')\right|$ "small" for all $j' \in [n]\backslash\{i'\}$ and $j \in [n]\backslash\{i\}$. Accordingly, the partial matching π_{ℓ} will be an improvement over $\pi_{\ell-1}$ unless there are many (of order roughly $|\{v \in [n] : \pi_{\ell-1}(v) \neq \pi(v)\}|$) vertices of G or G' with a considerable proportion of neighbors wrongly matched by $\pi_{\ell-1}$. This, however, can be ruled out with a high probability. The basic principle can be formulated as follows. If I is any random subset of $[n]$ containing a vast majority of the vertices of G, then for any positive constant $c > 0$ the set $\{i \in [n] : |\mathcal{N}_G(i) \cap I^c| \geq cpn\}$ has cardinality at most $\frac{1}{4}|I^c|$ with high probability; see Sect. 7 for deails.

Moreover, observe that the expected time complexity of Algorithm 4 is $n^{2+o(1)}$. The n^2 part comes from the loop over $i \in [n]$ and the "if" statement which consists in

searching over $i' \in [n]$. All the other computations can be done in $n^{o(1)}$ time because the neighborhoods are of typical size $n^{o(1)}$. The following theorem provides guarantees on the performance of Algorithm 4.

Theorem 2.4 (Refining a partial matching). *For any $\varepsilon \in (0, 1]$, there exists $n_0 > 0$ and $\kappa \in (0, 1)$ depending on ε with the following property. Let G^π and G' be the two graphs given by the correlated Erdős–Rényi graph model defined in Sect. 1.1 with underlying matching $\pi : [n] \to [n]$ and parameters n, p, and α such that*

$$n \geq n_0, \qquad (1+\varepsilon)\log n \leq pn \leq \frac{\sqrt{n}}{4\log n}, \qquad \alpha \in (0, \varepsilon/4].$$

Given a random matching $\hat{\pi} : [n] \to [n]$ (possibly depending on G^π and G'), and with G^π, G', $\hat{\pi}$ as the input, let $\tilde{\pi} : [n] \to [n]$ be the output of Algorithm 4. Then we have

$$\mathbb{P}\big\{\tilde{\pi} = \pi\big\} \geq \mathbb{P}\big\{|\{i \in [n] : \hat{\pi}(i) \neq \pi(i)\}| \leq \kappa n\big\} - \exp(-\varepsilon pn/10).$$

The following corollary of the above theorems is our final result on exact recovery of the underlying matching.

Corollary 2.5 (Exact matching). *Fix a constant $\varepsilon \in (0, 1]$. There exists a constant $n_0 > 0$ depending on ε and absolute constants $\alpha_0, R > 0$ with the following property. Let G^π and G' be the two graphs from the correlated Erdős–Rényi graph model defined in Sect. 1.1 with underlying matching $\pi : [n] \to [n]$ and parameters n, p, and α satisfying*

$$n \geq n_0, \qquad (1+\varepsilon)\log n \leq np \leq n^{\frac{1}{R\log\log n}}, \qquad 0 < \alpha < \alpha_0 \wedge (\varepsilon/4).$$

Run Algorithm 2 (with G^π and G' as input graphs) to obtain $B \in \{0, 1\}^{n\times n}$, then run Algorithm 3 to obtain $\hat{\pi} : [n] \to [n]$, and finally run Algorithm 4 to obtain $\tilde{\pi} : [n] \to [n]$. Then we have

$$\mathbb{P}\big\{\tilde{\pi} = \pi\big\} \geq 1 - n^{-10} - \exp(-\varepsilon pn/10).$$

Proof First, we apply Corollary 2.3 with $D = 10$ to obtain $\big|\big\{i \in [n] : \hat{\pi}(i) \neq \pi(i)\big\}\big| \leq 4n^{1-c}$ with probability at least $1 - n^{-10}$. Then, we choose κ depending on ε according to Theorem 2.4. If n is sufficiently large depending on ε, then $4n^{1-c} \leq \kappa n$, so Theorem 2.4 gives the result. □

Observe that Corollary 2.3 implies Theorem B from the introduction.

2.4 Further related work

The algorithms proposed above are related to several existing methods for graph matching. First, to achieve exact matching under the stronger condition $\alpha \leq (\log n)^{-C}$, the

paper [11] introduced a method based on comparing the degree profiles of vertices, that is, the empirical distributions of neighbors' degrees. The condition can be further improved to $\alpha \leq (\log\log n)^{-C}$ for sparse graphs by exploring neighbors' degree profiles. Note that the degree profiles of neighbors of a vertex are determined by the 3-neighborhood of the vertex. On the other hand, our algorithm uses degree information in a neighborhood of radius $\Theta(\log\log n)$ around each vertex, which is key to matching graphs with constant correlation. Prior to our work, large neighborhood statistics were used in the paper [21] which studied seeded graph matching — the version of the problem where a handful of correctly matched pairs of vertices are given to the algorithm as "seeds". We remark that, while the idea of leveraging degree statistics in neighborhoods is not new, our method of exploiting correlation via partition trees is novel.

The local tree structure in sparse graphs has been used in previous work for partial matching [15] and correlation detection [16]. It is worth noting that these papers considered local trees whose nodes are vertices of the observed graphs, while we consider partition trees whose nodes are *sets* of vertices. This is crucial to our analysis, which centers around estimating the overlaps between nodes of partition trees.

Moreover, the procedure of refining a partial matching (Algorithm 4) bears similarity to algorithms in prior works on similar topics [17, 19, 28]. The problems studied in these works, however, are inherently different from ours. To be more precise, it is assumed in these papers that the initial partial matching π_0 is *independent* from the observed graphs, while the initial matching we study can be an estimator computed from the observed graphs or even *adversarially* chosen. This adversarial setting has been studied in [2, 11], but they require a vanishing fraction of wrongly matched pairs in the initial matching, while our result can tolerate any constant fraction.

Finally, the strategy of matching vertices via comparing signature vectors appeared in the paper [20] by the current authors. In this previous paper, to exploit correlation between the two graphs, we started with comparing certain degree quantiles of the vertices and then refine the matched quantiles of vertices in two steps to obtain the final matching. The essence of this procedure is different from that in the current paper, although both eventually lead to weakly correlated signature vectors of length polylogarithmic in n. The more "global" strategy of comparing degree quantiles allows us to match $G(n, p)$ graphs with any average degree $np \geq (\log n)^C$ but requires a smaller noise level $\alpha \leq (\log\log n)^{-C}$. On the other hand, the current work adopts a more "local" approach of comparing neighborhoods that are typically trees, yielding a better condition $\alpha \leq$ const for sparse graphs.

3 Overview of the analysis

Our proof starts with some preparatory observations on the structure of large neighborhoods in sparse Erdős–Rényi Graphs, comprised in Sect. 4. The main goal is to show that with a high probability and for an appropriate choice of a parameter k, the k–neighborhoods of a vast majority of vertices are trees having sizes in a prescribed range, and with a prescribed statistics of nodes with very high or very low degrees.

The main tools in this section are standard concentration inequalities and standard properties of the binomial distribution.

The signature comparison is carried out in Sects. 5 and 6. We recall that the depth parameter m for constructing the signatures is double logarithmic in n. Section 5 starts with a crucial observation that the classes $T_s^m(i, G)$ and $T_s^m(i, G')$ from the partition trees of i in G and G', respectively, typically have intersections which introduce a detectable correlation of the degree statistics of their neighbors. More specifically, we show that under an appropriate graphs density assumption, with high probability almost every vertex i has $|T_s^m(i, G) \cap T_s^m(i, G')| \geq (np/2)^m(1-\upsilon)^m$, where $\upsilon > 0$ is an arbitrarily small constant. The proof is accomplished by induction, by considering $T_s^\ell(i, G) \cap T_s^\ell(i, G')$ for $0 \leq \ell \leq m$ (see Proposition 5.1). Note that the size of the intersection $|T_s^m(i, G) \cap T_s^m(i, G')|$ is still vanishing compared to the typical order of magnitude of $|T_s^m(i, G)|$ and $|T_s^m(i, G')|$ (that is, roughly $(np/2)^m$).

Further, in Sect. 5.2 we discuss the sparsification procedure which greatly simplifies the signature comparison. Sparsification is introduced to avoid the situation when for many pairs of distinct indices $s, s' \in \{-1, 1\}^m$, the sets $T_s^m(i, G)$ and $T_{s'}^m(i, G')$ still have a considerable intersection, which would introduce complex dependencies between components of the signature vectors of i in G and G'. While such an event appears difficult to control directly, by taking a relatively small uniform random subset J of indices in $\{-1, 1\}^m$ instead of the entire index set, we can guarantee that the undesired situation does not occur with high probability. More precisely, we are able to show that under some additional technical assumptions, the sets $R_s(i) := \mathcal{N}_G\big(T_s^m(i, G)\big) \cap \mathcal{N}_{G'}\big(T_{J\setminus\{s\}}^m(i, G')\big) \cap \mathcal{S}_{G_0}(i, m+1)$ and $R'_s(i) := \mathcal{N}_{G'}\big(T_s^m(i, G')\big) \cap \mathcal{N}_G\big(T_{J\setminus\{s\}}^m(i, G)\big) \cap \mathcal{S}_{G_0}(i, m+1)$, with $s \in J$, have small cardinalities with a high probability. The comparison of signatures of correct pairs of vertices is then accomplished in Sect. 5.3, with Sect. 5.4 summarizing the results.

Section 6, where the comparison of signatures of wrong pairs of vertices is carried out, has the same high-level structure as Sect. 5, although in that case somewhat more delicate estimates are required to show that the signature vectors of distinct vertices of G and G' are "essentially uncorrelated".

Construction of an exact matching between vertices of the two graphs from the approximate matching obtained in Theorem 2.1 and Corollary 2.3, is accomplished in Sect. 7. The goal of this section is to show that with high probability, Algorithm 4 described in Sect. 2, with the input partial matching given by Corollary 2.3, will produce the exact matching between G^π and G'.

4 Large neighborhoods in an Erdős–Rényi graph

In this section, we consider structural properties of vertex neighborhoods in sparse Erdős–Rényi graphs. While some of the statements (in particular, regarding the sets T_s^l from Algorithm 1) are method-specific, others, dealing with degree concentration and existence of cycles, are standard. Nevertheless, we prefer to provide the proofs for completeness. For the reader's convenience, we recall Chernoff's inequality for sums of Bernoulli random variables:

Lemma 4.1 (Chernoff's inequality; see, for example, [25, Section 2.3]). *Let $b_1, \dots, b_u$ be independent Bernoulli random variables with a parameter $q \in (0,1)$. Then*

$$\mathbb{P}\left\{\sum_{i=1}^{u} b_i > s\right\} \le \exp(-qu)\left(\frac{e\,qu}{s}\right)^s, \quad s > qu,$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{u} b_i < s\right\} \le \exp(-qu)\left(\frac{e\,qu}{s}\right)^s, \quad 0 < s < qu.$$

In particular, for every $s \ge e^2 qu$,

$$\mathbb{P}\left\{\sum_{i=1}^{u} b_i > s\right\} < \exp(-qu)\exp(-s),$$

and for every $s \in (qu, 2qu]$,

$$\mathbb{P}\left\{\sum_{i=1}^{u} b_i > s\right\} < \exp\left(-\frac{c(s-qu)^2}{qu}\right),$$

where $c > 0$ is a universal constant.

4.1 Cardinality estimates for large neighborhoods and vertex classes

Lemma 4.2 (Sizes of neighborhoods and their intersections). *For any $D > 1$, there is $K > 0$ and $n_0 \in \mathbb{N}$ depending only on D such that the following holds. Let G be a $G(n,p)$ graph with $n \ge n_0$ and $pn \ge \log n$. With probability at least $1 - n^{-D}$, we have that*

$$|\mathcal{B}_G(i,l)| \le K(np)^l \quad \text{for any } i, l \in [n]. \tag{1}$$

On the event that (1) holds, for any $m \in \mathbb{N}$ and any $i, j \in [n]$ such that $i \ne j$ and $G(\mathcal{B}_G(i,3m))$ is a tree, if $d := \operatorname{dist}_G(i,j) \le 2m$, then

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

Proof The first part of the lemma is standard: It is not hard to see that $|\mathcal{S}_G(i,l)|$ is stochastically dominated by a Binomial(n^l, p^l) random variable, so the bound follows from Chernoff's inequality (third estimate in Lemma 4.1) and a union bound. We omit the details.

For the second part, assume that (1) holds. Fix distinct $i, j \in [n]$ and let $d = \operatorname{dist}_G(i,j) \le 2m$. Then there exists a path γ from i to j of length d. Fix $v \in \mathcal{B}_G(i,m) \cap \mathcal{B}_G(j,m)$. Then there is a path from i to v in $\mathcal{B}_G(i,m)$ which we denote by γ_i, and

there is a path from j to v in $\mathcal{B}_G(j,m)$ which we denote by γ_j. Since $\mathcal{B}_G(i,m) \cup \mathcal{B}_G(j,m)$ is contained in $\mathcal{B}_G(i,3m)$ which is assumed to be a tree, the paths γ, γ_i, and γ_j are unique. Since the union of γ, γ_i, and γ_j is a tree, all of them must pass through a unique common vertex w. Since the length of γ_i is at most m, we have $v \in \mathcal{B}_G(w, m - \mathrm{dist}_G(i,w))$; similarly, we get $v \in \mathcal{B}_G(w, m - \mathrm{dist}_G(j,w))$. As a result, if $l := m - \max(\mathrm{dist}_G(w,i), \mathrm{dist}_G(w,j))$, then $v \in \mathcal{B}_G(w,l)$. Note that $m - d \le l \le m - \lceil d/2 \rceil$ in particular.

We now count the total number of vertices $v \in \mathcal{B}_G(i,m) \cap \mathcal{B}_G(j,m)$ by reversing the above reasoning: For any fixed w on the path γ, there are at most $K(np)^l$ vertices in $\mathcal{B}_G(w,l)$; this gives an upper bound on the number of possible v that connects to w. Letting w vary on the path γ and recalling the definition of l, we easily see that each $l = m - d, \ldots, m - \lceil d/2 \rceil$ corresponds to at most two w. Therefore, the total number of v can be bounded by

$$2 \sum_{l=m-d}^{m-\lceil d/2 \rceil} K(np)^l \le 4K(np)^{m-\lceil d/2 \rceil},$$

which finishes the proof (up to a change of the constant K). □

Lemma 4.3 (Sizes of vertex classes). *For any $C, D > 1$, there exists $Q > 0$, $R > 0$, and $n_0 \in \mathbb{N}$ depending only on C and D such that the following holds. Let G be a $G(n,p)$ graph with $n \ge n_0$ and*

$$\log n \le np \le n^{\frac{1}{R \log\log n}} \tag{2}$$

Fix a positive integer $m \le C \log\log n$. For $l \in [m]$, $s \in \{-1,1\}^l$, and $i \in [n]$, let $T_s^l(i)$ denote the class T_s^l of vertices constructed in `VertexSignature`(G,i,m) *(Algorithm 1). Then, with probability at least $1 - n^{-D}$, for any $i \in [n]$ such that $G(\mathcal{B}_G(i,m+1))$ is a tree, any $l \in [m]$, and any $s \in \{-1,1\}^l$, we have*

$$|T_s^l(i)| \le Q \left(\frac{np}{2}\right)^l.$$

Proof Fix $i \in [n]$ throughout the proof. We omit the subscript G in notations $\mathcal{B}_G, \mathcal{S}_G$, etc. for simplicity. For readability, we split the proof into a few parts.

Redefining the classes of vertices Let us define a modified version of the degree of a vertex and the classes $T_s^l = T_s^l(i)$ which coincide with the original definitions when $G(\mathcal{B}(i,m+1))$ is a tree. Namely, for $j \in \mathcal{S}(i,l)$, we set

$$\widehat{\deg}(j) := |\mathcal{N}(j) \cap \mathcal{S}(i,l+1)| + 1. \tag{3}$$

Define classes $\widehat{T}_s^l$ the same way as T_s^l but with $\deg(j)$ replaced by $\widehat{\deg}(j)$; that is, $\widehat{T}_\varnothing^0 = \{i\}$, and for $l \in \mathbb{N}_0$ and $s \in \{-1,1\}^l$,

$$\widehat{T}^{l+1}_{(s,+1)} = \left\{ j \in \mathcal{N}(\widehat{T}^l_s) \cap \mathcal{S}(i, l+1) : \widehat{\deg}(j) \geq np \right\},$$
$$\widehat{T}^{l+1}_{(s,-1)} = \left\{ j \in \mathcal{N}(\widehat{T}^l_s) \cap \mathcal{S}(i, l+1) : \widehat{\deg}(j) < np \right\}.$$

Note that if $G(\mathcal{B}(i, m+1))$ is a tree, then $\deg(j) = \widehat{\deg}(j)$ for any $j \in \mathcal{B}(i, m) \backslash \{i\}$, and thus $\widehat{T}^l_s = T^l_s$ for any $l \leq m$ and any $s \in \{-1, 1\}^l$. Therefore, it suffices to bound the cardinality of $\widehat{T}^l_s$ for $l \in [m]$.

Setup of the induction For the base case $l = 1$ of the induction, it is standard to show that

$$\mathbb{P}\left\{ |\widehat{T}^1_1| \leq Knp \right\} \geq \mathbb{P}\left\{ \widehat{\deg}(i) \leq Knp \right\} \geq 1 - n^{-D-1} \tag{4}$$

for a constant $K > 0$ depending only on D, since $np \geq \log n$ and $n \geq n_0 = n_0(D)$. The same bound holds for $|\widehat{T}^1_{-1}|$ as well.

Next, we set up the induction from $l \in [m-1]$ to $l+1$. Define

$$Q_l := 2K \left(1 + \frac{1}{m}\right)^l, \tag{5}$$

so $Q_l \leq 2eK$ for all $l \in [m]$. Let $s = (s_1, \ldots s_l)$ be such that $s_l = 1$. The case $s_l = -1$ can be handled in the same way. As the induction hypothesis, we assume that

$$|\widehat{T}^l_s| \leq Q_l (np/2)^l.$$

The induction step consists in proving that

$$|\widehat{T}^{l+1}_{(s,1)}| \leq Q_{l+1} (np/2)^{l+1}$$

with overwhelming probability (which will be explained more precisely below). The same bound for $\widehat{T}^{l+1}_{(s,-1)}$ can be established similarly, finishing the induction.

To make the high-probability statement in the induction precise, we define an event

$$\mathcal{E}_l := \left\{ \widehat{\deg}(j) \leq Knp \ \forall\, j \in \mathcal{B}(i, l) \right\} \tag{6}$$

for any $l \in [m]$. Moreover, let $\mathbb{P}_l$ and $\mathbb{E}_l$ denote respectively the probability and the expectation conditional on the subgraph $G(\mathcal{B}(i, l))$. For the induction step, we will show that conditional on an instance of $G(\mathcal{B}(i, l))$ such that $\mathcal{E}_{l-1}$ occurs, we have

$$\mathbb{P}_l \left\{ |\widehat{T}^{l+1}_{(s,1)}| \leq Q_{l+1} \left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \,\middle|\, |\widehat{T}^l_s| \leq Q_l \left(\frac{np}{2}\right)^l \right\} \geq 1 - n^{-D-1}. \tag{7}$$

Taking a union bound over (4) and (7) for $l \in [m-1]$ and $s \in \{-1,1\}^l$, we can remove the conditioning and obtain

$$\mathbb{P}\left\{|\widehat{T}_s^l| \le 2eK\left(\frac{np}{2}\right)^l \ \forall\, l \in [m],\ s \in \{-1,1\}^l\right\} \ge 1 - \sum_{l=1}^{m} 2^l \cdot n^{-D-1} \ge 1 - n^{-D}.$$

The induction step, part 1 We will estimate the size of the set $\mathcal{N}(\widehat{T}_s^l) \cap \mathcal{S}(i, l+1)$ first. Let us condition on an instance of $G(\mathcal{B}(i,l))$ such that $\mathcal{E}_{l-1}$ occurs. Further, fix a set $F \subset \mathcal{S}(i,l)$ such that $1 \le |F| \le Q_l(np/2)^l$, and condition on $\widehat{T}_s^l = F$. Then the random variables $\widehat{\deg}(j) - 1$ are independent across different j: The independence is ensured by the modified definition (3) of the degree of a vertex as we exclude the edges connecting different vertices in F.

Moreover, for any $j \in F$, $\widehat{\deg}(j) - 1$ is distributed as a Binomial$(n - |\mathcal{B}(i,l)|, p)$ random variable Z_j conditioned on $Z_j - 1 \ge np$. Since $\mathcal{E}_{l-1}$ occurs, we have $|\mathcal{B}(i,l)| \le 2(Knp)^l \le n^{0.1}$ in view of the conditions $l \le m \le C \log\log n$ and (2). It is not hard to see that $\mathbb{P}\{Z_j - 1 > np\} > 1/4$. Then Chernoff's inequality (fourth estimate in Lemma 4.1) yields

$$\begin{aligned}\mathbb{P}_l\left\{\widehat{\deg}(j) > np(1+t) \,\middle|\, \widehat{T}_s^l = F\right\} &= \mathbb{P}\left\{Z_j > np(1+t) \mid Z_j \ge np\right\} \\ &\le 4\,\mathbb{P}\left\{Z_j > np(1+t)\right\} \\ &\le 4\exp(-ct^2 np) \end{aligned} \tag{8}$$

for any $t \in (0,1)$ and a universal constant $c > 0$. In addition, a bound similar to (8) for $j \in \mathcal{S}(i,l)\backslash F$ can also be established, yielding

$$\mathbb{P}_l\left\{\max_{j \in \mathcal{S}(i,l)} \widehat{\deg}(j) > Knp \,\middle|\, \widehat{T}_s^l = F\right\} \le n^{-D-2}$$

for $K > 0$ depending only on D. Since we already condition on an instance of $G(\mathcal{B}(i,l))$ such that $\mathcal{E}_{l-1}$ occurs, the above inequality together with the definition of $\mathcal{E}_l$ in (6) implies that

$$\mathbb{P}_l\left\{\mathcal{E}_l \text{ occurs} \,\middle|\, \widehat{T}_s^l = F\right\} \ge 1 - n^{-D-2}. \tag{9}$$

Next, define an indicator

$$I_j := \mathbb{1}\left\{\widehat{\deg}(j) > np\left(1 + \frac{1}{2m}\right)\right\}.$$

Then (8) with $t = 1/(2m)$ shows that I_j is a Bernoulli(θ) random variable with $\theta < 4\exp(\frac{-cnp}{m^2})$. As $|F| \le Q_l(np/2)^l$, using the conditional independence of $\widehat{\deg}(j)$ for different j and applying Chernoff's inequality (first estimate in Lemma 4.1), we get

$$\mathbb{P}_l\left\{\sum_{j\in F} I_j \geq \frac{1}{2m}\left(\frac{np}{2}\right)^l \;\middle|\; \widehat{T}_s^l = F\right\} \leq \exp\left(-\frac{c'}{m^3}\left(\frac{np}{2}\right)^{l+1}\right) \leq n^{-D-2} \tag{10}$$

for a constant $c' > 0$, since $m \leq C \log\log n$, $np \geq \log n$, and $n \geq n_0 = n_0(C, D)$.

If $\sum_{j\in F} I_j < \frac{1}{2m}\left(\frac{np}{2}\right)^l$, the event $\mathcal{E}_l$ occurs, and $|F| \leq Q_l\left(\frac{np}{2}\right)^l$, then

$$\begin{aligned}\sum_{j\in F} \widehat{\deg}(j) &\leq \sum_{j\in F} I_j \cdot Knp + |F| \cdot np\left(1+\frac{1}{2m}\right)\\ &\leq \frac{1}{2m}\left(\frac{np}{2}\right)^l Knp + Q_l\left(\frac{np}{2}\right)^l np\left(1+\frac{1}{2m}\right)\\ &\leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1}\end{aligned}$$

by the definition of Q_l in (5). Combining this with (9) and (10) yields

$$\begin{aligned}&\mathbb{P}_l\left\{|\mathcal{N}(F)\cap\mathcal{S}(i,l+1)| \leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \;\middle|\; \widehat{T}_s^l = F\right\}\\ &\geq \mathbb{P}_l\left\{\sum_{j\in F}\widehat{\deg}(j) \leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \;\middle|\; \widehat{T}_s^l = F\right\}\\ &\geq 1-2n^{-D-2}\end{aligned}$$

for all F such that $|F| \leq Q_l\left(\frac{np}{2}\right)^l$. Removing the conditioning on $\widehat{T}_s^l = F$, we derive that

$$\begin{aligned}&\mathbb{P}_l\left\{|\mathcal{N}(\widehat{T}_s^l)\cap\mathcal{S}(i,l+1)| \leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \;\middle|\; |\widehat{T}_s^l| \leq Q_l\left(\frac{np}{2}\right)^l\right\}\\ &= \sum_{|F|\leq Q_l\left(\frac{np}{2}\right)^l} \mathbb{P}_l\left\{|\mathcal{N}(F)\cap\mathcal{S}(i,l+1)| \leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \;\middle|\; \widehat{T}_s^l = F\right\}\\ &\quad\cdot \mathbb{P}_l\left\{\widehat{T}_s^l = F \;\middle|\; |\widehat{T}_s^l| \leq Q_l\left(\frac{np}{2}\right)^l\right\}\\ &\geq 1-2n^{-D-2}.\end{aligned} \tag{11}$$

The induction step, part 2 For brevity, denote $H = \mathcal{N}\left(\widehat{T}_s^l\right)\cap\mathcal{S}(i,l+1)$. We further condition on an instance of $G(\mathcal{B}(i,l+1))$ such that

$$|H| \leq 2Q_l\left(1+\frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs},$$

that is, the high-probability event in (11) occurs. Then the quantities $\widehat{\deg}(j) - 1$ for $j \in H$ are i.i.d. Binomial$(n - |\mathcal{B}(i,l+1)|, p)$ random variables. For $j \in H$, denote

by Y_j the indicator of the event $\widehat{\deg}(j) \geq np$. Since $|\mathcal{B}(i, l+1)| \leq 2(Knp)^{l+1} \leq n^{0.1}$ by the conditions $l \leq m \leq C \log\log n$ and (2), we have

$$\left|\mathbb{E}_{l+1}\left[Y_j\right] - \frac{1}{2}\right| = \left|\mathbb{P}_{l+1}\left\{\widehat{\deg}(j) \geq np\right\} - \frac{1}{2}\right| \leq \frac{C_2 n^{0.1} p}{\sqrt{np}} \leq n^{-0.4}$$

where $C_2 > 0$ is a universal constant. Therefore,

$$|\widehat{T}^{l+1}_{(s,1)}| = \sum_{j\in H} Y_j \leq \sum_{j\in H}\left(Y_j - \mathbb{E}_{l+1}Y_j\right) + \left(\frac{1}{2} + n^{-0.4}\right)|H|.$$

Hoeffding's inequality then yields

$$\begin{aligned}
&\mathbb{P}_{l+1}\left\{|\widehat{T}^{l+1}_{(s,1)}| \geq \left(\frac{1}{2} + n^{-0.4}\right)|H| + \frac{1}{8m} Q_l \left(\frac{np}{2}\right)^{l+1}\right\} \\
&\leq \mathbb{P}_{l+1}\left\{\sum_{j\in H}\left(Y_j - \mathbb{E}_{l+1}Y_j\right) \geq \frac{1}{8m} Q_l \left(\frac{np}{2}\right)^{l+1}\right\} \\
&\leq \exp\left(-\frac{c}{|H|}\left(\frac{1}{8m} Q_l \left(\frac{np}{2}\right)^{l+1}\right)^2\right) \leq n^{-D-2}
\end{aligned}$$

if $m \leq C \log\log n$ and $n \geq n_0 = n_0(C, D)$, where we used $|H| \leq 2Q_l\left(1 + \frac{3}{4m}\right)(np/2)^{l+1}$. Moreover,

$$\begin{aligned}
&\left(\frac{1}{2} + n^{-0.4}\right)|H| + \frac{1}{8m} Q_l \left(\frac{np}{2}\right)^{l+1} \\
&\leq Q_l\left(1 + \frac{7}{8m}\right)\left(\frac{np}{2}\right)^{l+1} + \frac{1}{8m} Q_l \left(\frac{np}{2}\right)^{l+1} \\
&\leq Q_{l+1}\left(\frac{np}{2}\right)^{l+1},
\end{aligned}$$

so we obtain

$$\mathbb{P}_{l+1}\left\{|\widehat{T}^{l+1}_{(s,1)}| \leq Q_{l+1}\left(\frac{np}{2}\right)^{l+1}\right\} \geq 1 - n^{-D-2}.$$

It follows from the above inequality and (11) that

$$\begin{aligned}
&\mathbb{P}_l\left\{|\widehat{T}^{l+1}_{(s,1)}| \leq Q_{l+1}\left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \;\middle|\; |\widehat{T}^{l}_{s}| \leq Q_l\left(\frac{np}{2}\right)^{l}\right\} \\
&\geq \mathbb{P}_l\left\{|\widehat{T}^{l+1}_{(s,1)}| \leq Q_{l+1}\left(\frac{np}{2}\right)^{l+1} \;\middle|\; |H| \leq 2Q_l\left(1 + \frac{3}{4m}\right)\left(\frac{np}{2}\right)^{l+1},\right. \\
&\qquad \left.\mathcal{E}_l \text{ occurs, and } |\widehat{T}^{l}_{s}| \leq Q_l\left(\frac{np}{2}\right)^{l}\right\}
\end{aligned}$$

$$\cdot \mathbb{P}_l \left\{ |H| \le 2Q_l \left(1 + \frac{3}{4m}\right) \left(\frac{np}{2}\right)^{l+1} \text{ and } \mathcal{E}_l \text{ occurs} \,\middle|\, |\widehat{T}_s^l| \le Q_l \left(\frac{np}{2}\right)^l \right\}$$
$$\ge 1 - n^{-D-1},$$

which completes the induction. □

4.2 One-neighborhoods of typical vertices

Lemma 4.4 (Minimal degree of a typical vertex). *For any $\delta' > 0$ there are $n_0' \in \mathbb{N}$ and $c' > 0$ depending on δ' with the following property. Assume that $n > n_0'$ and that $r \in (0, 1)$ satisfies $rn \ge \log n$. Let Γ be a $G(n, r)$ Erdős–Rényi graph, and let J be any fixed subset of $[n]$. Then for any integer k such that $\max(2, |J| \exp(-c'rn)) \le k \le c'n$,*

$$\mathbb{P}\left\{\deg_\Gamma(i) < (1-\delta')rn \text{ for at least } k \text{ vertices } i \in J\right\} \le \exp(-c'rnk).$$

Proof For any distinct $i, j \in [n]$, let b_{ij} be the indicator of the event $\{(i, j) \text{ is an edge of } \Gamma\}$. Fix for a moment any $2 \le k \le |J|$ and k distinct vertices $v_1, v_2, \ldots, v_k$ in J. Let $1 \le \ell \le k$. Conditioned on any realization of random variables $(b_{v_i\, j})_{i \in [\ell-1],\, j \ne v_i}$, the variables $(b_{v_\ell,\, j})_{j \ne v_1, \ldots, v_\ell}$ are (conditionally) i.i.d. Bernoulli(r). Hence, by Chernoff's inequality (second estimate in Lemma 4.1),

$$\begin{aligned}
&\mathbb{P}\left\{ \sum_{j \ne v_1, \ldots, v_\ell} b_{v_\ell,\, j} < (1-\delta')rn \;\middle|\; (b_{v_i\, j})_{i \in [\ell-1],\, j \ne v_i} \right\} \\
&\quad \le \exp(-(n-\ell)r) \left(\frac{e(n-\ell)r}{(1-\delta')rn} \right)^{(1-\delta')rn} \\
&\quad \le \exp(-\delta' rn/2)
\end{aligned}$$

if $\ell \le k \le \delta' n/2$. This implies

$$\mathbb{P}\left\{ \sum_{j \ne v_1, \ldots, v_\ell} b_{v_\ell,\, j} < (1-\delta')rn \text{ for all } 1 \le \ell \le k \right\} \le \exp(-k\delta' rn/2).$$

By the above bound and a union bound over all possible subsets of J of cardinality k, we have

$$\begin{aligned}
\mathbb{P}\left\{\deg_\Gamma(i) < (1-\delta')rn \text{ for at least } k \text{ vertices } i \in J\right\} &\le \exp(-k\delta' rn/2) \binom{|J|}{k} \\
&\le \exp(-k\delta' rn/4)
\end{aligned}$$

if $k \ge e|J| \exp(-\delta' rn/4)$. □

An immediate consequence of the above lemma is the following result.

Lemma 4.5 (Degrees of typical vertices). *For any $\kappa \in (0, 1/2)$, there exists $c > 0$ and $n_0 \in \mathbb{N}$ depending on κ with the following property. Suppose that $n > n_0$ and $nr \geq \log n$. Let Γ be a $G(n, r)$ Erdős–Rényi graph. Then with probability at least $1 - \exp(-crn \log n)$,*

$$\left|\left\{i \in [n] : \ |\mathcal{N}_\Gamma(i)| \geq (1-\kappa)nr\right\}\right| > n - n^{1-c}. \tag{12}$$

Proof We apply Lemma 4.4 with $J = [n]$ to obtain the following: There is a constant $c > 0$ depending on κ such that for any integer k such that $\max(2, n\exp(-cnr)) \leq k \leq cn$,

$$\mathbb{P}\{|\mathcal{N}_\Gamma(i)| < (1-\kappa)nr \text{ for at least } k \text{ vertices } i \in [n]\} \leq \exp(-ckrn).$$

Taking

$$k := \left\lceil \max\left(\log n, \ n\exp(-cnr)\right) \right\rceil \leq n^{1-c},$$

we see that $|\mathcal{N}_\Gamma(i)| \geq (1-\kappa)nr$ for more than $n - n^{1-c}$ vertices $i \in [n]$ with probability at least $1 - \exp(-crn \log n)$. □

Lemma 4.6 (Number of common neighbors). *For any $\delta'' > 0$ there are $n_0'' \in \mathbb{N}$ and $c'' > 0$ depending on δ'' with the following property. Assume that $n > n_0''$ and that $r \in (0, 1/2]$ satisfies $rn \geq \log n$. Let Γ be a $G(n, r)$ Erdős–Rényi graph, and let J be any fixed subset of $[n]$ such that $|J| \cdot rn \log n \leq \sqrt{n}$. Then for every $v \in J$ we have*

$$\mathbb{P}\left\{\left|\mathcal{N}_\Gamma(v) \cap \left(\mathcal{N}_\Gamma(J\backslash\{v\}) \cup J\right)\right| \geq \delta''rn\right\} \leq \exp(-c''rn \log n).$$

Proof For any distinct $i, j \in [n]$, let b_{ij} denote the indicator of the event $\{(i, j) \text{ is an edge of } \Gamma\}$. For any given vertex i of Γ, we have, by Chernoff's inequality (third estimate in Lemma 4.1),

$$|\mathcal{N}_\Gamma(i)| \leq ern \log n$$

with probability at least $1 - \exp(-rn)\exp(-ern \log n)$, whence the event

$$\mathcal{E}' := \left\{|\mathcal{N}_\Gamma(J\backslash\{v\})| \leq |J| \cdot ern \log n\right\}$$

has probability at least $1 - \exp(-ern \log n)$. Observe that $\mathcal{N}_\Gamma(J\backslash\{v\})$ (and so $\mathcal{E}'$) is measurable w.r.t. the collection of variables $(b_{ij})_{i\in J\backslash\{v\},\, j\neq i}$. Conditioned on any realization of $(b_{ij})_{i\in J\backslash\{v\},\, j\neq i}$ such that $\mathcal{E}'$ holds, the variables b_{vj}, $j \in \mathcal{N}_\Gamma(J\backslash\{v\})\backslash J$, are (conditionally) i.i.d. Bernoulli(r). Hence, by Chernoff's inequality (first estimate in Lemma 4.1) and the assumption $|J| \cdot rn \log n \leq \sqrt{n}$,

$$\mathbb{P}\left\{\left|\mathcal{N}_\Gamma(v) \cap \left(\mathcal{N}_\Gamma(J\backslash\{v\})\backslash J\right)\right| \geq \delta''rn/2 \mid \mathcal{E}'\right\} \leq \left(\frac{e|J| \cdot er^2 n \log n}{\delta''rn/2}\right)^{\delta''rn/2}$$

$$\leq \exp(-c_1 rn \log n)$$

for some $c_1 > 0$ depending on δ''. Finally, we note that, again by Chernoff's inequality and the assumption on $|J|$,

$$\mathbb{P}\left\{|\mathcal{N}_\Gamma(v) \cap J| \geq \delta'' rn/2 \mid \mathcal{E}'\right\} \leq \exp(-c_2 rn \log n)$$

for some $c_2 > 0$ depending on δ'', and the result follows. □

Lemma 4.7 (Neighbors' degrees in a typical graph). *For any $\kappa, \lambda \in (0, 1/2)$, there exist $c, \delta > 0$ and $n_0 \in \mathbb{N}$ depending on κ and λ with the following property. Let $n > n_0$ and $rn \geq \log n$, and let Γ be a $G(n, r)$ Erdős–Rényi graph. Further, let J be a fixed subset of $[n]$ such that*

$$\frac{1}{2}\log n \leq |J| \leq \frac{\sqrt{n}}{rn \log n}. \tag{13}$$

Then with probability at least $1 - \exp(-crn \log n)$,

$$\left|\left\{j \in \mathcal{N}_\Gamma(i) : \deg_\Gamma(j) \geq nr + \delta\sqrt{nr}\right\}\right| \geq (1/2 - \kappa)nr, \quad \text{and} \tag{14a}$$

$$\left|\left\{j \in \mathcal{N}_\Gamma(i) : \deg_\Gamma(j) \leq nr - \delta\sqrt{nr}\right\}\right| \geq (1/2 - \kappa)nr \tag{14b}$$

for at least $|J| - \max(\lambda \log n, |J| \exp(-crn))$ vertices $i \in J$.

Proof Choose $\delta' = \delta'' := \kappa/4$, and let c', c'' be the corresponding constants from Lemmas 4.4 and 4.6. Denote by $\mathcal{E}$ the event

$$\begin{aligned}\mathcal{E} := \Big\{&|\mathcal{N}_\Gamma(v)| \geq (1-\delta')rn \text{ for at least } |J| - \max\big((\lambda/2)\log n,\\ &|J|\exp(-c'rn)\big) \text{ vertices } v \in J,\\ &|\mathcal{N}_\Gamma(v) \cap (\mathcal{N}_\Gamma(J\backslash\{v\}) \cup J)| < \delta'' rn \text{ for all } v \in J, \text{ and}\\ &|\mathcal{N}_\Gamma(J)| \leq |J| \cdot ern \log n\Big\}.\end{aligned}$$

In view of Lemmas 4.4 and 4.6, and by applying Chernoff's inequality (Lemma 4.1) to estimate the upper tails of $|\mathcal{N}_\Gamma(v)|$ for $v \in J$, it is not difficult to see that the event $\mathcal{E}$ has probability at least

$$\begin{aligned}&1 - \exp(-c'(\lambda/2)\, rn \log n) - n \cdot \exp(-c'' rn \log n) - \exp(-ern \log n)\\ &\geq 1 - \exp(-2c_1 rn \log n)\end{aligned}$$

for a constant $c_1 > 0$ depending on κ and λ.

Note that $\mathcal{E}$ is measurable with respect to the variables $(b_{ij})_{i\in J,\, j\neq i}$, where b_{ij} is the indicator of the event $\{(i, j) \text{ is an edge of } \Gamma\}$. Condition on any realization of the variables $(b_{ij})_{i\in J,\, j\neq i}$ such that the event $\mathcal{E}$ holds. Denote $U_i := \mathcal{N}_\Gamma(i)\backslash(\mathcal{N}_\Gamma(J\backslash\{i\}) \cup$

J) for $i \in J$. Observe that the sets U_i, $i \in J$, are disjoint, and everywhere on $\mathcal{E}$ we have

$$\left|\left\{i \in J : |U_i| \geq (1-\kappa/2)rn\right\}\right| \geq |J| - \max\left((\lambda/2)\log n, |J|\exp(-c'rn)\right). \quad (15)$$

Let $\delta \in (0,1)$ be a parameter to be chosen later. Fix $i \in J$ and let

$$Z_j := \sum_{k \in [n]\setminus(J \cup \mathcal{N}_\Gamma(J))} b_{jk}, \quad j \in U_i.$$

Then the indicators of the events

$$\left\{Z_j \geq nr + \delta\sqrt{nr}\right\}, \quad j \in U_i,$$

are conditionally i.i.d. given $(b_{vj})_{v \in J,\, j \neq v}$. Observe that for every $j \in U_i$, the variable Z_j is conditionally Binomial($n - |J \cup \mathcal{N}_\Gamma(J)|$,$r$), and so

$$nr + \delta\sqrt{nr} - \mathbb{E}[Z_j] \leq 2\delta\sqrt{\mathrm{Var}(Z_j)} + 3r^{1/2}\sqrt{\mathrm{Var}(Z_j)}$$

in view of the assumption (13) on $|J|$. Hence, by the Berry–Esseen theorem (or by a direct computation), there is a choice of δ, as a function of κ, so that

$$Z_j \geq nr + \delta\sqrt{nr}$$

with conitional probability at least $1/2 - \kappa/4$. Thus, applying Hoeffding's inequality, for every i such that $|U_i| \geq (1-\kappa/2)rn$, we have

$$\left|\left\{j \in U_i : \ Z_j \geq nr + \delta\sqrt{nr}\right\}\right| = \sum_{j \in U_i} \mathbb{1}\left\{Z_j \geq nr + \delta\sqrt{nr}\right\} \geq (1/2 - \kappa)rn$$

with conditional probability at least $1 - \exp(-c_2\kappa^2 rn)$, for some universal constant $c_2 > 0$.

Finally, the variables

$$\left|\left\{j \in U_i : \sum_{k \in [n]\setminus(J \cup \mathcal{N}_\Gamma(J))} b_{jk} \geq nr + \delta\sqrt{nr}\right\}\right|, \quad i \in J,$$

are conditionally mutually independent given $(b_{vj})_{v \in J,\, j \neq v}$. It remains to apply Chernoff's inequality (second estimate in Lemma 4.1) to the sum of the indicators

$$\sum_{i \in J:\, |U_i| \geq (1-\kappa/2)rn} \mathbb{1}\left\{\left|\left\{j \in U_i : \sum_{k \in [n]\setminus(J \cup \mathcal{N}_\Gamma(J))} b_{jk} \geq nr + \delta\sqrt{nr}\right\}\right| \geq (1/2 - \kappa)rn\right\}$$

to obtain the following: Since the above sum is over at least $|J| - \max\left((\lambda/2)\log n, |J|\exp(-c'rn)\right)$ summands by (15), with probability at least $1 - \exp(-c_3 rn\log n)$,

$$\left|\left\{j \in U_i : \sum_{k\in[n]\setminus(J\cup\mathcal{N}_\Gamma(J))} b_{jk} \geq nr + \delta\sqrt{nr}\right\}\right| \geq (1/2-\kappa)rn$$

for at least $|J| - \max\left((3\lambda/4)\log n, |J|\exp(-c_3 rn)\right)$ vertices $i \in J$, where $c_3 > 0$ is a constant depending on κ and λ. Then (14a) follows immediately. The condition (14b) is verified by analogy. □

We now prove a lemma similar to the above, but with $J = [n]$.

Lemma 4.8 (Neighbors' degrees of typical vertices). *For any $\kappa \in (0, 1/2)$, there exist $c, \delta > 0$ and $n_0 \in \mathbb{N}$ depending on κ with the following property. Suppose that $n > n_0$ and $\log n \leq rn \leq n^{1/4}$. Let Γ be a $G(n, r)$ Erdős–Rényi graph. Then with probability at least $1 - \exp(-crn\log n)$,*

$$\left|\left\{j \in \mathcal{N}_\Gamma(i) : \deg_\Gamma(j) \geq nr + \delta\sqrt{nr}\right\}\right| \geq (1/2-\kappa)nr, \quad \text{and} \tag{16a}$$

$$\left|\left\{j \in \mathcal{N}_\Gamma(i) : \deg_\Gamma(j) \leq nr - \delta\sqrt{nr}\right\}\right| \geq (1/2-\kappa)nr. \tag{16b}$$

for at least $n - n^{1-c}$ vertices $i \in [n]$.

Proof Let $\ell := \lceil rn^{3/2}\log n\rceil$. We partition $[n]$ into ℓ fixed subsets $J_1, \ldots, J_\ell$ such that

$$|J_i| \leq n/\ell \leq \frac{\sqrt{n}}{rn\log n}$$

for every $i \in [\ell]$. By Lemma 4.7 and a union bound over $J_1, \ldots, J_\ell$, the probability that

$$\left|\left\{j \in \mathcal{N}_\Gamma(i) : \deg_\Gamma(j) \geq nr + \delta\sqrt{nr}\right\}\right| < (1/2-\kappa)nr \tag{17}$$

for at most $\max(\log n, |J_\ell|\exp(-crn))$ vertices $i \in J_\ell$ for all $\ell \in [\ell]$ is at least

$$1 - \ell\exp(-4crn\log n) \geq 1 - \exp(-2crn\log n),$$

where $c > 0$ depends on κ. On this high-probability event, (17) holds for at most

$$\ell\cdot\max(\log n,\ |J_\ell|\exp(-crn)) \leq \max(2rn^{3/2}\log^2 n,\ n\exp(-crn)) \leq n^{1-2c}$$

vertices $i \in [n]$, since $\log n \leq rn \leq n^{1/4}$. This proves (16a). The proof of (16b) is analogous. □

4.3 Tree structure of large neighborhoods

Lemma 4.9 (Counting subsets via probabilistic method). *Let $n, d, k \in \mathbb{N}$ be such that $2k \le d \le n$. Fix an integer $M \ge (2n/d)^k d \log(en/d)$. There exists a collection of M subsets $A_i \subset [n]$ with $|A_i| = k$ for each $i \in [M]$ such that the following holds. For every subset $B \subset [n]$ with $|B| = d$, there is $i \in [M]$ such that $A_i \subset B$.*

Proof Fix a subset $B \subset [n]$ with $|B| = d$. Let $A_1, \dots, A_M$ be i.i.d. uniformly random subsets of $[n]$ with $|A_i| = k$ for each $i \in [M]$. Then we have

$$\mathbb{P}\{A_i \subset B\} = \frac{d(d-1)\cdots(d-k+1)}{n(n-1)\cdots(n-k+1)} > \Big(\frac{d}{2n}\Big)^k$$

and thus

$$\mathbb{P}\{A_i \not\subset B \ \forall\, i \in [M]\} < \bigg(1 - \Big(\frac{d}{2n}\Big)^k\bigg)^M < \exp\bigg(-M\Big(\frac{d}{2n}\Big)^k\bigg).$$

A union bound over all subsets $B \subset [n]$ with $|B| = d$ yields

$$\mathbb{P}\{\exists\, B \subset [n],\ |B| = d,\ \text{s.t. } A_i \not\subset B\ \forall\, i \in [M]\} < \exp\bigg(-M\Big(\frac{d}{2n}\Big)^k\bigg) \cdot \binom{n}{d} \le 1$$

if $M \ge (2n/d)^k d \log(en/d)$. Taking the complement of the above event completes the proof. □

Lemma 4.10 (Tree structure in a typical graph). *For $n \in \mathbb{N}$ and $r \in (0,1)$, let Γ be a $G(n,r)$ Erdős–Rényi graph. For any $x \in \mathbb{N}$, the probability that there are at least $(5x)^3(\log n)^6(nr)^{3x}$ vertices of Γ whose x–neighborhoods are not trees, is bounded from above by $\exp(-\log^2 n)$.*

Proof For an integer $k \ge 2$, fix for a moment k distinct vertices $v_1, \dots, v_k \in [n]$, and consider the event $\mathcal{E}$ that the x–neighborhood of each of the vertices is *not* a tree. Given any realization of Γ from $\mathcal{E}$, we shall construct an auxiliary subgraph H of Γ iteratively as follows.

At the beginning of the process, H is empty. Since the x–neighborhood of v_1 contains a cycle, there is a cycle in this neighborhood which we denote as an ordered collection of vertices $S_1 = (S_1(u))_{u=0}^{z_1}$, satisfying $S_1(0) = S_1(z_1)$ and $3 \le z_1 \le 2x$, and a path $P_1 = (P_1(u))_{u=0}^{\ell_1}$ of length $0 \le \ell_1 \le x$ starting at v_1 and ending at $S_1(0)$, with $P_1 \cup S_1$ containing only one cycle, that is, S_1. We then add the vertex and the edge sets of $P_1 \cup S_1$ to H.

Next, we consider the vertex v_2. One of the following two assertions is then true for v_2:

- Either there is a cycle $S_2 = (S_2(u))_{u=0}^{z_2}$ in Γ and a path $P_2 = (P_2(u))_{u=0}^{\ell_2}$ starting at v_2 and ending at $S_2(0)$, with the same conditions on lengths and cycle number as above, and such that the vertex set of $P_2 \cup S_2$ is *disjoint* from the vertex set of $P_1 \cup S_1$. In this case, we add $P_2 \cup S_2$ to H.

- Or, there is a path $P_2 = (P_2(u))_{u=0}^{\ell_2}$ of length $\ell_2 \le x$, starting at v_2 and ending at some vertex of $P_1 \cup S_1$, such that $P_1 \cup S_1 \cup P_2$ contains only one cycle, S_1. We then add P_2 to H.

To summarize the first two steps informally, the subgraph H now consists of either (1) two connected components, with one component containing a cycle S_1 and a path P_1 connecting v_1 to a vertex in S_1, and the other component containing a cycle S_2 and a path P_2 connecting v_2 to a vertex in S_2; or (2) one connected component, containing a cycle S_1, a path P_1 connecting v_1 to a vertex in S_1, and a path P_2 connecting v_2 to a vertex in $S_1 \cup P_1$.

Next, we continue to do the same construction for v_3, after which the possibilities of the subgraph H include the following: (1) three components, $S_1 \cup P_1$, $S_2 \cup P_2$, and $S_3 \cup P_3$; (2) two components, $S_1 \cup P_1 \cup P_2$ and $S_3 \cup P_3$; (3) two components, $S_1 \cup P_1$ and $S_2 \cup P_2 \cup P_3$; (4) one component, $S_1 \cup P_1 \cup P_2 \cup P_3$.

More rigorously, repeating the process for all $v_1, \dots, v_k$, we can guarantee that there exists an integer $1 \le h \le k$, a partition of $[k]$ into h non-empty subsets $T_1, \dots, T_h$, cycles $S_j = (S_j(u))_{u=0}^{z_j}$, $1 \le j \le h$, with $3 \le z_j \le 2x$, and paths $P_i = (P_i(u))_{u=0}^{\ell_i}$, $1 \le i \le k$, with $0 \le \ell_i \le x$, such that all of the following conditions are satisfied:

- each of $T_1, \dots, T_h$ consists of consecutive integers, and they form an ordered partition of $[k]$ (for example, $k = 6$, $h = 3$, $T_1 = \{1, 2, 3\}$, $T_2 = \{4\}$, and $T_3 = \{5, 6\}$);
- $3 \le z_j \le 2x$ for all $1 \le j \le h$, and $0 \le \ell_i \le x$ for all $1 \le i \le k$;
- $S_j(0) = S_j(z_j)$ for each $1 \le j \le h$;
- $P_i(0) = v_i$ for all $1 \le i \le k$ (ith path starts at v_i);
- $P_i(\ell_i) = S_j(0)$ for all $1 \le j \le h$ and $i = \min T_j := \min\{i' : i' \in T_j\}$ (the path P_i attaches vertex v_i to the cycle S_j at $S_j(0)$ for $i = \min T_j$);
- $P_i(\ell_i) \in \bigcup_{i' \in T_j,\, i' < i} P_{i'} \cup S_j$, for all $1 \le j \le h$ and $i \in T_j \backslash \{\min T_j\}$ (each vertex $v_i \in T_j$ is attached via the path P_i to the union of S_j and the paths $P_{i'}$ with $i' \in T_j$ and $i' < i$);
- The vertices $S_j(u)$, $0 \le u < z_j$, $1 \le j \le h$, and $P_i(u)$, $0 \le u < \ell_i$, $1 \le i \le k$, are all distinct;
- All unordered pairs of vertices $\{S_j(u), S_j(u+1)\}$, $0 \le u < z_j$, $1 \le j \le h$, and $\{P_i(u), P_i(u+1)\}$, $0 \le u < \ell_i$, $1 \le i \le k$, are edges of Γ;
- As a consequence of the above conditions, the subgraphs $\bigcup_{i' \in T_j} P_{i'} \cup S_j$ are disjoint across different $1 \le j \le h$;
- The subgraph $H := \left(\bigcup_{j=1}^{h} S_j\right) \cup \left(\bigcup_{i=1}^{k} P_i\right)$ contains exactly h cycles and h connected components (one cycle in each component).

Thus, the probability of $\mathcal{E}$ can be (crudely) estimated from above as

$$\sum_{h=1}^{k} \sum_{T_1 \sqcup \cdots \sqcup T_h = [k]} \sum_{z_1,\ldots,z_h \in \{3,\ldots,2x\}} \sum_{\ell_1,\ldots,\ell_k \in \{0,\ldots,x\}} r^{z_1+\cdots+z_h+\ell_1+\cdots+\ell_k} n^{z_1+\cdots+z_h+\ell_1+\cdots+\ell_k-k} (3xk)^k,$$

where the exponent of r is the number of edges, the exponent of n is the number of vertices (besides the fixed $v_1, \ldots, v_k$), and the factor $(3xk)^k$ is bounding the number of possible vertices to which $v_1, \ldots, v_k$ are attached via the paths. To bound the above sum, we note that

- the number of all possible ordered partitions $T_1, \ldots, T_h$ of $[k]$ is at most $\binom{2k}{k}$ by a standard "stars and bars" argument;
- the sums over $z_1, \ldots, z_h$ and $\ell_1, \ldots, \ell_k$ give a factor at most $(2x)^k (x+1)^k$;
- $z_1 + \cdots + z_h + \ell_1 + \cdots + \ell_k \le 3kx$.

Combining the above estimates, we conclude that the probability of $\mathcal{E}$ is bounded from above by

$$\binom{2k}{k} (2x)^k (x+1)^k (nr)^{3kx} (3xk)^k n^{-k} \le (3kx)^{3k} (nr)^{3kx} n^{-k}.$$

Finally, let T be the collection of all vertices of Γ whose x–neighborhoods are not trees. We are interested in bounding $\mathbb{P}\{|T| \ge d\}$ where $d \ge 2k$ using the above bound on the probability of $\mathcal{E}$. By Lemma 4.9, it suffices to take a union bound over $\lceil (2n/d)^k d \log(en/d) \rceil$ subsets of cardinality k, yielding

$$\mathbb{P}\{|T| \ge d\} \le (3kx)^{3k} (nr)^{3kx} n^{-k} \lceil (2n/d)^k d \log(en/d) \rceil \le \exp(-\log^2 n)$$

once we take $k = \lceil \log^2 n \rceil$ and $d \ge (5x)^3 (\log n)^6 (nr)^{3x}$. □

4.4 Typical vertices in the parent graph

Recall that the parent graph G_0 is a $G(n, q)$ random graph. We consider parameters n, q, and $m \in \mathbb{N}$ satisfying

$$n \ge n_0, \qquad \log n \le nq \le n^{\frac{1}{R \log\log n}}, \qquad m \le C \log\log n,$$

where n_0, R, and C are positive constants.

Let us first define a *typical* vertex of the parent graph G_0. The definition incorporates all the structural properties of vertex neighborhoods that will be important for signature comparison.

Definition 4.11 We say that a vertex $i \in [n]$ of G_0 is *typical* with parameters $m \in \mathbb{N}$, $\kappa > 0$, $K > 1$, and $\delta > 0$, and write $i \in \mathbf{Typ}_{G_0}(m, \kappa, K, \delta)$, if i has the following properties:

(A1) $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree.

(A2) $\deg_{G_0}(j) \le Knq$ for any $j \in \mathcal{B}_{G_0}(i, m+1)$.

(A3) $|\mathcal{S}_{G_0}(i,l)| > (1-\kappa)nq \cdot 3^{l-1}$ and $|\mathcal{B}_{G_0}(i,l)| \le K(nq)^l$ for all $l \in [m]$.

(A4) For any $l \in \{0, \dots, m\}$, the following holds. Denote by $\mathrm{HD}(l)$ the set of vertices in the l-sphere whose degrees are relatively high:

$$\mathrm{HD}(l) := \{j \in \mathcal{S}_{G_0}(i,l) : \deg_{G_0}(j) > (1-\kappa)nq\}.$$

Then

$$|\mathcal{S}_{G_0}(i,l) \backslash \mathrm{HD}(l)| \le \frac{\kappa}{K \cdot 3^l} |\mathcal{S}_{G_0}(i,l)|.$$

(A5) For any $l \in \{0, \dots, m-1\}$, the following holds. For $j \in \mathcal{S}_{G_0}(i,l)$, denote by $V_+(j)$ the set of its neighbors whose degrees are noticeably larger than the mean:

$$V_+(j) := \left\{j' \in \mathcal{N}_{G_0}(j) \cap \mathcal{S}_{G_0}(i, l+1) : \deg_{G_0}(j') > nq + \delta\sqrt{nq}\right\}.$$

Furthermore, denote by $W_+(l)$ the subset of $\mathcal{S}_{G_0}(i,l)$ for which the sets $V_+(j)$ are large:

$$W_+(l) := \left\{j \in \mathcal{S}_{G_0}(i,l) : |V_+(j)| \ge (1/2 - \kappa)nq\right\}.$$

Then

$$|\mathcal{S}_{G_0}(i,l) \backslash W_+(l)| \le \frac{\kappa}{K \cdot 3^l} |\mathcal{S}_{G_0}(i,l)|.$$

(A6) For any $l \in \{0, \dots, m-1\}$, the following holds. For $j \in \mathcal{S}_{G_0}(i,l)$, denote by $V_-(j)$ the set of its neighbors whose degrees are noticeably smaller than the mean:

$$V_-(j) := \left\{j' \in \mathcal{N}_{G_0}(j) \cap \mathcal{S}_{G_0}(i, l+1) : \deg_{G_0}(j') < nq - \delta\sqrt{nq}\right\}.$$

Furthermore, let

$$W_-(l) := \left\{j \in \mathcal{S}_{G_0}(i,l) : |V_-(j)| \ge (1/2 - \kappa)nq\right\}.$$

Then

$$|\mathcal{S}_{G_0}(i,l) \backslash W_-(l)| \le \frac{\kappa}{K \cdot 3^l} |\mathcal{S}_{G_0}(i,l)|.$$

The sets $\mathrm{HD}(l)$, $W_+(l)$, $W_-(l)$ depend, of course, on i. However, we do not mention this dependence explicitly to lighten the notation. The following is the main statement of the subsection.

Proposition 4.12 *For any $\kappa \in (0, 1/2)$ and $C, D > 1$, there exist $\delta, c \in (0, 1/2)$ and $K, R, n_0 > 1$ depending on κ, C, D such that the following holds. If*

$$n \geq n_0, \qquad \log n \leq nq \leq n^{\frac{1}{R \log\log n}}, \qquad m \leq C \log\log n,$$

then with high probability, most vertices of a $G(n, q)$ graph G_0 are typical with parameters m, κ, K, and δ:

$$\mathbb{P}\left\{ \left| \mathbf{Typ}_{G_0}(m, \kappa, K, \delta) \right| \geq n - n^{1-c} \right\} \geq 1 - n^{-D}.$$

Proof Consider each condition in Definition 4.11:

1. By Lemma 4.10, with probability at least $1 - \exp(-\log^2 n)$, there are at most

$$(5m + 5)^3 (\log n)^6 (nq)^{3m+3} \leq \sqrt{n}$$

vertices of G_0 whose $(m + 1)$–neighborhoods are not trees, where the above inequality holds because $m \leq C \log\log n$ and $nq \leq n^{\frac{1}{R \log\log n}}$ for R depending on C.
2. By Chernoff's inequality (third estimate in Lemma 4.1) and the condition $nq \geq \log n$, we in fact have $\deg_{G_0}(j) \leq Knq$ for all $j \in [n]$ with probability at least $1 - n^{-D-1}$, where K depends on D.
3. The upper bound $|\mathcal{B}_{G_0}(i, l)| \leq K(np)^l$ holds simultaneously for all $i \in [n]$ with probability at least $1 - n^{-D-1}$ by (1).
   For the lower bound in the case $l = 1$, the bound (12) in Lemma 4.5 with $\Gamma = G_0$ shows that $|\mathcal{S}_{G_0}(i, 1)| = |\mathcal{N}_{G_0}(i)| \geq (1 - \kappa)nq$ for at least $n - n^{1-c_1}$ vertices $i \in [n]$ with probability at least $1 - n^{-D-1}$, where $c_1 > 0$ depends on κ and D.
   The lower bound in the case $l \in \{2, \ldots, m\}$ is trivial, because the size of $\mathcal{S}_{G_0}(i, l)$ is of order $(nq)^l$ which is much larger than $nq \cdot 3^{l-1}$ for all $i \in [n]$ with overwhelming probability.
4. For $l = 0$, if i satisfies Condition (A3), then $\deg(i) > (1-\kappa)nq$ and so $\mathcal{S}_{G_0}(i, 0) = \mathrm{HD}(0) = \{i\}$.
   Next, fix $l \in [m]$. We will show that with high probability, any vertex $i \in [n]$ satisfying Conditions (A1) and (A3) also satisfies Condition (A4). Towards this end, condition on any realization of the subgraph $G_0(\mathcal{B}_{G_0}(i, l-1))$ and all the edges between $\mathcal{B}_{G_0}(i, l-1)$ and its complement. Let $\tilde{\mathbb{P}}$ denote the conditional probability. Under this conditioning, $\mathcal{S}_{G_0}(i, l)$ is determined. Let $\widehat{G} := G_0\big([n] \backslash \mathcal{B}_{G_0}(i, l-1)\big)$ and $\widehat{n} := n - |\mathcal{B}_{G_0}(i, l-1)|$. Then $\widehat{G}$ is conditionally a $G(\widehat{n}, q)$ graph and $\mathcal{S}_{G_0}(i, l)$ is a fixed subset of vertices.
   For $j \in \mathcal{S}_{G_0}(i, l)$, let us define

$$\widehat{\mathrm{HD}}(l) := \{ j \in \mathcal{S}_{G_0}(i, l) : \deg_{\widehat{G}}(j) + 1 > (1 - \kappa)nq \}.$$

If $G_0(\mathcal{B}_{G_0}(i, m+1))$ is a tree, then $\deg_{\widehat{G}}(j) + 1 = \deg_{G_0}(j)$, and so $\widehat{\mathrm{HD}}(l) = \mathrm{HD}(l)$. Applying Lemma 4.4 with $\Gamma = \widehat{G}$ and $J = \mathcal{S}_{G_0}(i, l)$, we obtain

$$\tilde{\mathbb{P}}\left\{\left|\mathcal{S}_{G_0}(i,l)\backslash\widehat{\mathrm{HD}}(l)\right| < k\right\} \geq 1 - \exp(-c_2 q\widehat{n}k) \tag{18}$$

for a constant $c_2 > 0$ depending on κ, where $\max(2, |J|\exp(-c_2 q\widehat{n})) \leq k \leq c_2\widehat{n}$. If i satisfies Condition (A3), then

$$|\mathcal{B}_{G_0}(i,l)| \leq K(nq)^m \leq Kn^{\frac{C}{R}},$$

since $m \leq C\log\log n$ and $nq \leq n^{\frac{1}{R\log\log n}}$ where R depends on C and D. As a result, $0.9n \leq \widehat{n} \leq n$ and

$$|J|\exp(-c_2 q\widehat{n}) \leq Kn^{\frac{C}{R}}\exp(-c_2 q\widehat{n}) \leq 1,$$

where the last inequality holds because $nq \geq \log n$ and $n \geq n_0 = n_0(\kappa, C, D)$ if we choose $R = R(C)$ appropriately. Therefore, we can take $k = \left\lceil \frac{2(D+2)}{c_2} \right\rceil$ so that the error probability in (18) is at most n^{-D-2}. Finally, since

$$k = \left\lceil \frac{2(D+2)}{c_2} \right\rceil < \frac{\kappa}{K\cdot 3^l}\frac{nq\cdot 3^{l-1}}{2} < \frac{\kappa}{K\cdot 3^l}|\mathcal{S}_{G_0}(i,l)|$$

for i satisfying Condition (A3), we obtain from (18) that with probability at least $1 - n^{-D-1}$, for any vertex $i \in [n]$ satisfying Conditions (A1) and (A3),

$$|\mathcal{S}_{G_0}(i,l)\backslash\mathrm{HD}(l)| < \frac{\kappa}{K\cdot 3^l}|\mathcal{S}_{G_0}(i,l)|.$$

5. For $l = 0$, the bound (16a) in Lemma 4.8 with $\Gamma = G_0$ shows that with probability at least $1 - n^{-D-2}$, we have $\mathcal{S}_{G_0}(i, 0) = \{i\} = W_+(0)$ for at least $n - n^{1-c_3}$ vertices $i \in [n]$, where $c_3 > 0$ depends on κ and D. Therefore, Condition (A5) holds for these vertices.
   Next, fix $l \in [m-1]$. We will show that with high probability, any vertex $i \in [n]$ satisfying Conditions (A1) and (A3) also satisfies Condition (A5). As in the proof of Condition (A4), we let $\tilde{\mathbb{P}}$ denote the probability conditional on any realization of the subgraph $G_0(\mathcal{B}_{G_0}(i, l-1))$ and all the edges between $\mathcal{B}_{G_0}(i, l-1)$ and its complement. Again, let $\widehat{G} := G_0\big([n]\backslash\mathcal{B}_{G_0}(i, l-1)\big)$ and $\widehat{n} := n - |\mathcal{B}_{G_0}(i, l-1)|$ so that $\widehat{G}$ is conditionally a $G(\widehat{n}, q)$ graph.
   For $j \in \mathcal{S}_{G_0}(i, l)$, let

$$\widehat{V}_+(j) := \left\{j' \in \mathcal{N}_{\widehat{G}}(j) :\ \deg_{\widehat{G}}(j') > nq + \delta\sqrt{nq}\right\}$$

and

$$\widehat{W}_+(l) := \left\{j \in \mathcal{S}_{G_0}(i,l) :\ |\widehat{V}_+(j)| \geq (1/2 - \kappa)nq\right\}.$$

If $G_0(\mathcal{B}_{G_0}(i,m))$ is a tree, then it is not hard to see that $\widehat{V}_+(j) = V_+(j)$ and thus $\widehat{W}_+(l) = W_+(l)$. To bound $|\mathcal{S}_{G_0}(i,l)\backslash\widehat{W}_+(l)|$, we apply Lemma 4.7 with $\Gamma = \widehat{G}$, $J = \mathcal{S}_{G_0}(i,l)$, and $\lambda = \frac{\kappa}{4K}$. Note that by Condition (A3) together with the conditions $\log n \le nq \le n^{\frac{1}{R\log\log n}}$ and $m \le C\log\log n$, we have

$$\frac{1}{2}\log n \le |\mathcal{S}_{G_0}(i,l)| \le \frac{\sqrt{n}}{rn\log n},$$

so Lemma 4.7 can indeed be applied. Hence, we obtain that with probability at least $1-\exp(-c_4 rn\,\log n) \ge 1-n^{-D-2}$,

$$|\mathcal{S}_{G_0}(i,l)\backslash\widehat{W}_+(l)| \le \max\left(\frac{\kappa}{4K}\log n,\ |\mathcal{S}_{G_0}(i,l)|\exp(-c_4 rn)\right) \le \frac{\kappa}{K\cdot 3^l}|\mathcal{S}_{G_0}(i,l)|$$

by Condition (A3) and $m \le C\log\log n$, where $c_4, \delta > 0$ depend on κ and D.

6. This part is analogous to the previous one.

Finally, by a union bound, with probability at least $1-n^{-D}$, the number of non-typical vertices is at most $\sqrt{n} + n^{1-c_1} + 2n^{1-c_3} \le n^{1-c}$ for a constant $c > 0$. The proof is therefore complete. □

5 Signatures of correct pairs of vertices

The goal of this section is to show that with probability close to one, for almost every vertex i, the signatures $f(i)$ and $f'(i)$ computed in the graphs G and G' are close to each other, in the sense that the sparsified ℓ_2–distance of appropriately normalized signatures is less than a given threshold. The constant correlation between G and G' introduces a very large noise so that the normalized signatures are still "almost orthogonal", and a high precision of the estimates is required to distinguish this case from the case when signatures of different vertices are compared.

5.1 Overlap between neighborhoods of a typical vertex in the child graphs

Let the graphs G_0, G, and G' be given by the correlated Erdős–Rényi graph model with parameters n, p, α as defined in Sect. 1.1. That is, G_0 is a $G(n,q)$ Erdős–Rényi graph, where $q := \frac{p}{1-\alpha}$. Conditional on G_0, the subgraph G is obtained by removing every edge of G_0 independently with probability α, and G' is a conditionally i.i.d. copy of G. Fix $m \in \mathbb{N}$. For $l \in [m]$, $s \in \{-1,1\}^l$, and $i \in [n]$, let $T_s^l(i,G)$ denote the set T_s^l of vertices constructed in `VertexSignature`(G,i,m) (Algorithm 1); similarly, let $T_s^l(i,G')$ and $T_s^l(i,G_0)$ denote the sets constructed by Algorithm 1 when the input graph is G' and G_0 respectively. Recall that collections $\mathbf{Typ}_{G_0}(m,\kappa,K,\delta)$ of vertices of the parent graph G_0 were described in Definition 4.11. In what follows, $\mathbb{P}_0$ denotes the probability conditional on an instance of G_0.

Proposition 5.1 *For any parameters $\kappa, K, \delta, D > 0$, there exist $\alpha_0 \in (0,1)$ and $n_0 > 0$ depending only on these parameters such that the following holds for any*

$\alpha \in (0, \alpha_0)$ and $n \geq n_0$ satisfying $nq \geq \log n$. Condition on an instance of the parent graph G_0. Fix $m \in \mathbb{N}$. Then with (conditional) probability at least $1 - n^{-D}$, for every typical vertex $i \in \mathbf{Typ}_{G_0}(m, \kappa, K, \delta)$ and any $s \in \{-1, 1\}^m$,

$$|T_s^m(i, G) \cap T_s^m(i, G')| \geq (np/2)^m (1 - 8\kappa)^m.$$

Proof Fix a typical vertex $i \in \mathbf{Typ}_{G_0}(m, \kappa, K, \delta)$, and let $l \in \{0, \dots, m-1\}$. We start with proving the following claim showing that with overwhelming probability, the degrees of all vertices in $\mathrm{HD}(l)$ in the child graphs G and G' remain relatively large, where $\mathrm{HD}(l)$ is defined in Condition (A4).

Claim 5.2 *There exists $\alpha_0 > 0$ depending on κ, K, and D such that for any $\alpha \in (0, \alpha_0)$,*

$$\mathbb{P}_0\big\{ \deg_G(j) \wedge \deg_{G'}(j) > (1 - 2\kappa)np \ \forall\, j \in HD(l) \big\} \geq 1 - n^{-D-1}.$$

To prove the claim, let $j \in \mathrm{HD}(l)$. Then

$$\begin{aligned}
&\mathbb{P}_0\left\{\deg_G(j) \leq (1-2\kappa)np\right\} + \mathbb{P}_0\left\{\deg_{G'}(j) \leq (1-2\kappa)np\right\} \\
&\quad \leq \mathbb{P}_0\left\{(1-\alpha)\deg_{G_0}(j) - \deg_G(j) > \kappa np\right\} \\
&\qquad + \mathbb{P}_0\left\{(1-\alpha)\deg_{G_0}(j) - \deg_{G'}(j) > \kappa np\right\}.
\end{aligned}$$

Note that $\deg_G(j)$ and $\deg_{G'}(j)$ are Binomial$(\deg_{G_0}(j), 1-\alpha)$ random variables conditional on G_0, where $\deg_{G_0}(j) \leq Knq$ by Condition (A2). Hence, the right-hand side of the inequality above is bounded by $2\exp\big(-c\frac{\kappa^2 np}{\alpha}\big)$ for a constant $c > 0$ depending on K. We can choose α sufficiently small depending on κ, K, and D such that this error probability is at most n^{-D-2}. The claim follows by taking the union bound over $j \in \mathrm{HD}(l)$.

Next, we will show that with overwhelming probability, the sets $W_+(l)$ and $W_-(l)$ defined in Conditions (A5) and (A6) respectively remain at least the same size under a slight change of their definitions. Namely, to take advantage of the independence of degrees of different vertices, we consider the spheres around i and the neighborhoods in G_0, but evaluate the degrees in G and G'. Specifically, we prove the following claim.

Claim 5.3 *There exists $\alpha_0 > 0$ depending on κ, K, δ, and D such that for any $\alpha \in (0, \alpha_0)$, the following holds. For $j \in \mathcal{S}_{G_0}(i, l)$, denote by $V_+^{G,G'}(j)$ the set of its neighbors in G_0 whose degrees in G and G' are larger than the mean:*

$$V_+^{G,G'}(j) := \big\{ j' \in \mathcal{N}_{G_0}(j) \cap \mathcal{S}_{G_0}(i, l+1) :\ \deg_G(j') > np \text{ and } \deg_{G'}(j') > np \big\}.$$

Then

$$\mathbb{P}_0\left\{ |V_+^{G,G'}(j)| \geq (1/2 - 2\kappa)nq \ \forall\, j \in W_+(l) \right\} \geq 1 - n^{-D-1}.$$

To prove this claim, consider a vertex $j \in W_+(l)$ where $W_+(l)$ is defined in Condition (A5), and estimate the probability of the event that $|V_+^{G,G'}(j)| < (1/2 - 2\kappa)nq$. Note that since i is typical, these events are independent for different $j \in W_+(l)$ conditional on G_0. Consider $j' \in V_+(j)$, so that $\deg_{G_0}(j') > nq + \delta\sqrt{nq}$. The distribution of the random variable $\deg_G(j')$ is conditionally Binomial$(\deg_{G_0}(j'), 1-\alpha)$. Therefore, assuming $\alpha \le 1/2$, by Bernstein's inequality we have

$$\begin{aligned}\mathbb{P}_0\big\{\deg_G(j') \le nq(1-\alpha)\big\} &\le \mathbb{P}_0\big\{\deg_{G_0}(j') - \deg_G(j') \ge \alpha \deg_{G_0}(j') + (1-\alpha)\delta\sqrt{nq}\big\} \\ &\le \exp\bigg(-\frac{c(1-\alpha)^2\delta^2 nq}{\alpha \deg_{G_0}(j') + (1-\alpha)\delta\sqrt{nq}}\bigg),\end{aligned}$$

where $c > 0$ is a universal constant, and where $\deg_{G_0}(j') \le Knq$ by Condition (A2). Let $\tau > 0$ be a number depending on κ, K, and D which will be chosen soon. Then, since n is large, we can choose $\alpha_0 \in (0, 1/2]$ depending on δ, K, and τ (and thus only on κ, K, δ, and D) such that for any $\alpha \in (0, \alpha_0)$, the above error probability is less than $\tau/2$. Similarly, one can show that

$$\mathbb{P}_0\big\{\deg_{G'}(j') \le nq(1-\alpha)\big\} \le \tau/2,$$

and so

$$\begin{aligned}\mathbb{P}_0\big\{j' \in V_+(j)\backslash V_+^{G,G'}(j)\big\} &= \mathbb{P}_0\big\{\deg_G(j') \le nq(1-\alpha)\big\} \\ &\quad + \mathbb{P}_0\big\{\deg_{G'}(j') \le nq(1-\alpha)\big\} \le \tau.\end{aligned}$$

Note that $|V_+(j)| \le \deg_{G_0}(j) \le Knq$, and that the events $\{j' \in V_+(j)\backslash V_+^{G,G'}(j)\}$ are independent for different $j' \in V_+(j)$ conditional on G_0 for a typical vertex i. Assuming that $\tau K \le \kappa$, we get by Chernoff's inequality (first estimate in Lemma 4.1),

$$\mathbb{P}_0\big\{|V_+(j)\backslash V_+^{G,G'}(j)| \ge \kappa nq\big\} \le \bigg(\frac{e\tau Knq}{\kappa nq}\bigg)^{\kappa nq}.$$

As $nq \ge \log n$, we can choose τ depending only on κ, K, and D so that the bound above does not exceed n^{-D-2}. By the union bound,

$$\begin{aligned}&\mathbb{P}_0\big\{\exists\, j \in W_+(l) \text{ s.t. } |V_+^{G,G'}(j)| < (1/2 - 2\kappa)nq\big\} \\ &\quad \le \mathbb{P}_0\big\{\exists\, j \in W_+(l) \text{ s.t. } |V_+(j)\backslash V_+^{G,G'}(j)| \ge \kappa nq\big\} \le n^{-D-1},\end{aligned}$$

finishing the proof of the claim.

Applying the same argument, one can establish a similar claim for the set $W_-(l)$.

Claim 5.4 *There exists $\alpha_0 > 0$ depending on κ, K, δ, and D such that for any $\alpha \in (0, \alpha_0)$, the following holds. For $j \in \mathcal{S}_{G_0}(i,l)$, denote by $V_-^{G,G'}(j)$ the set of its neighbors in G_0 whose degrees in G and G' are smaller than the mean:*

$$V_-^{G,G'}(j) = \big\{j' \in \mathcal{N}_{G_0}(j) \cap \mathcal{S}_{G_0}(i, l+1) :\ \deg_G(j') < np \text{ and } \deg_{G'}(j') < np\big\}.$$

Then

$$\mathbb{P}_0\left\{|V_-^{G,G'}(j)| \geq (1/2-2\kappa)nq \;\forall\, j \in W_-(l)\right\} \geq 1-n^{-D-1}.$$

Equipped with Claims 5.2, 5.3, and 5.4, we can complete the proof of the proposition. Let $\mathcal{E}(i,l)$ be the event that the following statements hold:

- $\deg_G(j) > (1-2\kappa)np$ and $\deg_{G'}(j) > (1-2\kappa)np$ for at least $\left(1-\frac{\kappa}{K\cdot 3^l}\right)|\mathcal{S}_{G_0}(i,l)|$ vertices $j \in \mathcal{S}_{G_0}(i,l)$;
- $|V_+^{G,G'}(j)| \geq (1/2-2\kappa)nq$ and $|V_-^{G,G'}(j)| \geq (1/2-2\kappa)nq$ for at least $\left(1-\frac{\kappa}{K\cdot 3^l}\right)|\mathcal{S}_{G_0}(i,l)|$ vertices $j \in \mathcal{S}_{G_0}(i,l)$.

Then the above claims together with Conditions (A4), (A5), and (A6) imply that

$$\mathbb{P}_0\left\{\bigcap_{l=0}^{m-1}\mathcal{E}(i,l)\right\} \geq 1-n^{-D}.$$

Assuming that the event $\bigcap_{l=0}^{m-1}\mathcal{E}(i,l)$ occurs, we will show that

$$|T_s^l(i,G)\cap T_s^l(i,G')| \geq (1-8\kappa)^l(np/2)^l \tag{19}$$

for all $l \in \{0,\dots,m\}$ and $s \in \{-1,1\}^l$ by induction on l. For $l=0$, this inequality trivially holds. Assume that (19) holds for $l \in \{0,\dots,m-1\}$ and $s \in \{-1,1\}^l$, and consider, for instance, $s'=(s,1) \in \{-1,1\}^{l+1}$. The case $s'=(s,-1) \in \{-1,1\}^{l+1}$ is handled the same way.

Let W be the set of all vertices $j \in T_s^l(i,G)\cap T_s^l(i,G')$ such that $\deg_G(j) > (1-2\kappa)np$, $\deg_{G'}(j) > (1-2\kappa)np$, and $|V_+^{G,G'}(j)| \geq (1/2-2\kappa)nq$. Then

$$|W| \geq |T_s^l(i,G)\cap T_s^l(i,G')| - 2\frac{\kappa}{K\cdot 3^l}|\mathcal{S}_{G_0}(i,l)| \geq (1-2\kappa)|T_s^l(i,G)\cap T_s^l(i,G')|,$$

where the last inequality relies on the induction hypothesis (19) and that $|\mathcal{S}_{G_0}(i,l)| \leq K(nq)^l$ in Condition (A3).

For any $j \in W$, the entire set $V_+^{G,G'}(j)$ is contained in $T_{(s,1)}^{l+1}(i,G)\cap T_{(s,1)}^{l+1}(i,G')$. Since these sets are disjoint for different $j \in W$,

$$\begin{aligned}|T_{(s,1)}^{l+1}(i,G)\cap T_{(s,1)}^{l+1}(i,G')| &\geq \sum_{j\in W}|V_+^{G,G'}(j)| \geq |W|\cdot(1/2-2\kappa)nq\\ &\geq (1/2-4\kappa)nq\,|T_s^l(i,G)\cap T_s^l(i,G')|\\ &\geq (1-8\kappa)^{l+1}(np/2)^{l+1},\end{aligned}$$

where the last inequality follows from the induction hypothesis. □

For the rest of this section, we fix a positive integer m and simplify the notation of classes of vertices by omitting m as follows: For $s \in \{-1,1\}^l$ and $i \in [n]$, let $T_s(i)$

denote the set T_s^m of vertices constructed by `VertexSignature`(G, i, m) (Algorithm 1); similarly, let $T'_s(i)$ denote the set constructed by `VertexSignature`(G', i, m). Moreover, for any subset $J \subset \{-1, 1\}^m$, we define

$$T_J(i) := \bigcup_{s \in J} T_s(i) \quad \text{and} \quad T'_J(i) := \bigcup_{s \in J} T'_s(i).$$

5.2 Sparsification

For pairs of distinct indices $s, s' \in \{-1, 1\}^m$, the sets $T_s(i)$ and $T'_{s'}(i)$ may have a considerable intersection, which introduces complex dependencies between components of the signature vectors of i in G and G'. To tackle this issue, we now use sparsification—taking a small random subset J of indices in $\{-1, 1\}^m$ instead of the entire index set—to guarantee that the undesired situation does not occur for too many pairs $s, s' \in J$ with high probability. We first state a lemma from [20].

Lemma 5.5 (Lemma 16 of [20]). *Fix a constant $S > 0$ and an even integer $k \in \mathbb{N}$. Let Ω and Ω' be two finite sets, and let*

$$\Omega = \bigcup_{i=1}^{k} \Omega_i \quad \text{and} \quad \Omega' = \bigcup_{i=1}^{k} \Omega'_i$$

be partitions of Ω and Ω' respectively such that $|\Omega'_i| \le S/k$ for all $i \in [k]$. Furthermore, let $w \in \{2, 3, \ldots, k/2\}$ and let I be a uniform random subset of $[k]$ of cardinality $2w$. Then, for any $L \ge 1$ and $\rho \in (0, 1/4)$ such that ρw is an integer, we have

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

Proof This is a restatement of Lemma 16 of [20] with $\gamma^{-1}|\Omega'|$ replaced by S, and L replaced by L/γ; the same proof works to give the statement above. Note that in that lemma, it is assumed in addition that $|\Omega'_i| \ge \gamma|\Omega'|/k$, but this condition is never used in the proof. □

The following lemma shows that the intersection between the neighbors of $T_s(i)$ in G and the neighbors of $T'_{s'}(i)$ in G' is not too large for most pairs $s, s' \in J$. More precisely, in (22), we let $R_s(i)$ denote the intersection between the neighbors of $T_s(i)$ and the neighbors of $T'_{s'}(i)$ for any $s' \in J$, $s' \ne s$, and we define $R'_s(i)$ analogously. The lemma states that there is a subset $\tilde{J}(i) \subset J$ that is almost as large as J (note (20) and that we take $2w = |J| \ge 2(\log n)^4$ in Lemma 5.11) such that $R_s(i)$ and $R'_s(i)$ are sufficiently small for all $s \in \tilde{J}(i)$. Note the extra factor $1/2^m$ in (21a) for $s \in \tilde{J}(i)$ compared to (21b) for any $s \in J$. This will be crucial to controlling $\eta_s(i)$ and $\eta'_s(i)$ in Lemma 5.7 and subsequent estimates in Lemma 5.11.

Lemma 5.6 (Sparsification). *For constants $C_1, C_2 > 0$, there exists $K = K(C_1, C_2) > 0$ with the following property. Let J be a uniform random subset of $\{-1, 1\}^m$ of cardinality $2w$ for an integer $w > 2(\log n)^2$. With respect to the randomness of J, the following holds with probability at least $1 - \exp(-(\log n)^{1.5})$ for any vertex $i \in [n]$: If*

- *$G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree,*
- *$\deg_G(j) \vee \deg_{G'}(j) \le C_1 np$ for all $j \in \mathcal{B}_{G_0}(i, m)$, and*
- *$|T_s(i)| \vee |T'_s(i)| \le C_2\big(\frac{np}{2}\big)^m$ for all $s \in \{-1, 1\}^m$,*

then there is a subset $\tilde{J}(i) \subset J$ such that

$$|J \setminus \tilde{J}(i)| < (\log n)^2, \tag{20}$$

and

$$|R_s(i)| \vee |R'_s(i)| \le K w^4 \frac{(np)^{m+1}}{4^m} \quad \text{for all } s \in \tilde{J}(i), \tag{21a}$$

$$|R_s(i)| \vee |R'_s(i)| \le K \frac{(np)^{m+1}}{2^m} \quad \text{for all } s \in J, \tag{21b}$$

where

$$R_s(i) := \mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{N}_{G'}\big(T'_{J\setminus\{s\}}(i)\big) \cap \mathcal{S}_{G_0}(i, m+1), \tag{22a}$$

$$R'_s(i) := \mathcal{N}_{G'}\big(T'_s(i)\big) \cap \mathcal{N}_G\big(T_{J\setminus\{s\}}(i)\big) \cap \mathcal{S}_{G_0}(i, m+1). \tag{22b}$$

Proof Fix a vertex i such that the three conditions in the lemma hold. Since $G_0(\mathcal{B}_{G_0}(i, m+1))$ is a tree and by definition $\mathcal{S}_G(i, m) = \bigcup_{s\in\{-1,1\}^m} T_s(i)$ and $\mathcal{S}_{G'}(i, m) = \bigcup_{s\in\{-1,1\}^m} T'_s(i)$, we have partitions

$$\mathcal{S}_G(i, m+1) = \bigcup_{s\in\{-1,1\}^m} \Big(\mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\Big),$$

$$\mathcal{S}_{G'}(i, m+1) = \bigcup_{s\in\{-1,1\}^m} \Big(\mathcal{N}_{G'}\big(T'_s(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\Big).$$

Using the conditions $|T_s(i)| \vee |T'_s(i)| \le C_2\big(\frac{np}{2}\big)^m$ for all $s \in \{-1, 1\}^m$ and $\deg_G(j) \vee \deg_{G'}(j) \le C_1 np$ for all $j \in \mathcal{B}_{G_0}(i, m)$, we see that

$$\Big|\mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\Big| \vee \Big|\mathcal{N}_{G'}\big(T'_s(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\Big| \le C_1 C_2 \frac{(np)^{m+1}}{2^m}.$$

Note that (21b) is a consequence of the above bound.

Moreover, we can apply Lemma 5.5 with $\Omega = \mathcal{S}_G(i, m+1)$, $\Omega' = \mathcal{S}_{G'}(i, m+1)$, $k = 2^m$, $S = C_1 C_2 (np)^{m+1}$, $L = 8ew^3$, and $\rho = \frac{1}{2w}(\log n)^2$ to obtain the following:

With probability at least $1 - \exp(-(\log n)^2/2)$,

$$\left|\left\{s \in J : \exists t \in J\backslash\{s\} \text{ s.t. } \left|\mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{N}_{G'}\big(T'_t(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\right| \geq 8ew^3C_1C_2\frac{(np)^{m+1}}{4^m}\right\}\right| < \frac{1}{2}(\log n)^2. \tag{23}$$

A similar estimate holds if T and T' are swapped. Next, define

$$\tilde{J}(i) := \left\{s \in J : |R_s(i)| \vee |R'_s(i)| \leq 16ew^4C_1C_2\frac{(np)^{m+1}}{4^m}\right\}$$

which is a superset of

$$\left\{s \in J : \forall t \in J\backslash\{s\}, \left|\mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{N}_{G'}\big(T'_t(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\right| \leq 8ew^3C_1C_2\frac{(np)^{m+1}}{4^m}\right\}$$
$$\bigcap \left\{s \in J : \forall t \in J\backslash\{s\}, \left|\mathcal{N}_{G'}\big(T'_s(i)\big) \cap \mathcal{N}_G\big(T_t(i)\big) \cap \mathcal{S}_{G_0}(i, m+1)\right| \leq 8ew^3C_1C_2\frac{(np)^{m+1}}{4^m}\right\}.$$

As a result of (23) and the counterpart with T and T' swapped, we see that (20) holds. Moreover, (21a) holds by the definition of $\tilde{J}(i)$.

Finally, a union bound over vertices $i \in [n]$ completes the proof. □

5.3 Difference between signatures of a typical pair

Throughout this subsection, we fix a vertex $i \in [n]$, a subset $J \subset \{-1, 1\}^m$ of cardinality $2w$ for $w \in \mathbb{N}$, and a subset $\tilde{J}(i) \subset J$. Moreover, we condition on the neighborhoods $G_0(\mathcal{B}_{G_0}(i, m+1))$, $G(\mathcal{B}_{G_0}(i, m+1))$, and $G'(\mathcal{B}_{G_0}(i, m+1))$ such that the following statements hold for fixed constants $K, \kappa > 0$:

(B1) $G_0(\mathcal{B}_{G_0}(i, m+1))$ is a tree;
(B2) $|\mathcal{B}_{G_0}(i, m+1)| \leq n^{0.1}$;
(B3) $\deg_G(j) \vee \deg_{G'}(j) \leq Knp$ for all $j \in \mathcal{B}_{G_0}(i, m)$ for a constant $K > 0$;
(B4) $|T_s(i)| \vee |T'_s(i)| \leq K(np/2)^m$ for all $s \in \{-1, 1\}^m$;
(B5) (20), (21a), and (21b) hold;
(B6) $\left|\left\{j \in \mathcal{S}_{G_0}(i, m) : \deg_{G\cap G'}(j) \leq \frac{2}{3}np(1-\alpha)\right\}\right| \leq (nq/3)^m$, where $G \cap G'$ denotes the graph on $[n]$ whose edge set is the intersection of those of G and G';
(B7) $|T_s(i) \cap T'_s(i)| \geq (np/2)^m(1 - 8\kappa)^m$ for all $s \in \{-1, 1\}^m$.

We consider the randomness with respect to the remaining possible edges of the graphs; let $\tilde{\mathbb{P}}$ and $\tilde{\mathbb{E}}$ denote the conditional probability and expectation respectively. For any $j \in \mathcal{S}_{G_0}(i, m+1)$, it is not hard to see that the random variable $\deg_G(j) - 1$ is conditionally Binomial$(\tilde{n}, p)$ where

$$\tilde{n} := n - |\mathcal{B}_{G_0}(i, m+1)| \geq n - n^{0.1},$$

and these binomial variables are independent across different $j \in \mathcal{S}_{G_0}(i, m+1)$.

Lemma 5.7 (Small overlaps). *For any $D, K > 0$, there exists $K' = K'(D, K) > 0$ such that the following holds. Define $R_s(i)$ and $R'_s(i)$ as in (22). With (conditional) probability at least $1 - n^{-D}$,*

$$|\eta_s(i)| \vee |\eta'_s(i)| \le K' \frac{(np)^{m/2+1}}{2^m} w^2 \sqrt{\log n} \quad \text{for all } s \in \tilde{J}(i), \tag{24a}$$

$$|\eta_s(i)| \vee |\eta'_s(i)| \le K' \frac{(np)^{m/2+1}}{2^{m/2}} \sqrt{\log n} \quad \text{for all } s \in J, \tag{24b}$$

where

$$\eta_s(i) := \sum_{j \in R_s(i)} \big(\deg_G(j) - 1 - np\big) \quad \text{and} \quad \eta'_s(i) := \sum_{j \in R'_s(i)} \big(\deg_{G'}(j) - 1 - np\big). \tag{25}$$

Proof Recall that the variables $\deg_G(j) - 1$, $j \in R_s(i)$, are conditionally independent Binomial$(\tilde{n}, p)$ where $\tilde{n} \ge n - n^{0.1}$. As a result, $\sum_{j \in R_s(i)} \big(\deg_G(j) - 1\big) \sim$ Binomial$(\tilde{n}|R_s(i)|, p)$, and so

$$\begin{aligned} |\eta_s(i)| &= \Big| \sum_{j \in R_s(i)} \big(\deg_G(j) - 1 - np\big) \Big| \\ &\le K_2\big(\sqrt{np|R_s(i)| \log n} + \log n\big) + (n - \tilde{n}) p\, |R_s(i)| \\ &\le K_3 \sqrt{np(|R_s(i)| + 1) \log n} \end{aligned}$$

with probability at least $1 - n^{-D-1}$ for constants $K_2, K_3 > 0$ depending on D. We then combine the above bound with Condition (B5) to obtain (24a) and (24b). Finally, a union bound over $s \in \{-1, 1\}^m$ completes the proof. □

Lemma 5.8 (Correlated binomial). *Let $\tilde{\mathbb{P}}$ and $\tilde{\mathbb{E}}$ denote the conditional probability and expectation respectively, defined at the beginning of this subsection. Fix a subset $I \subset \mathcal{S}_{G_0}(i, m+1)$. Let*

$$A := \sum_{j \in I} \big(\deg_G(j) - \deg_{G'}(j)\big) \quad \text{and} \quad B := \sum_{j \in I} \big(\deg_G(j) - 1 - \tilde{n}p\big)$$

where $\tilde{n} = n - |\mathcal{B}_{G_0}(i, m+1)|$. Then we have

- $\tilde{\mathbb{E}}[A] = \tilde{\mathbb{E}}[B] = 0$;
- $\tilde{\mathbb{E}}[A^2] = 2\alpha p\, \tilde{n}|I|$ *and* $\tilde{\mathbb{E}}[B^2] = p(1-p)\tilde{n}|I|$;
- $\tilde{\mathbb{P}}\{|A| \ge t\} \le 2\exp\Big(\frac{-t^2/2}{\tilde{\mathbb{E}}[A^2] + t/3}\Big)$ *and* $\tilde{\mathbb{P}}\{|B| \ge t\} \le 2\exp\Big(\frac{-t^2/2}{\tilde{\mathbb{E}}[B^2] + t/3}\Big)$ *for every $t > 0$.*

Proof We first consider the variable B. Since $\sum_{j\in I}(\deg_G(j)-1)$ is conditionally Binomial$(\tilde{n}|I|, p)$, the mean, variance, and tail bound for B are all standard facts.

Next, consider the variable A. Note that we can write

$$A = \sum_{j\in I} \sum_{\ell\in[n]\setminus\mathcal{B}_{G_0}(i,m+1)} X_{j\ell}(Y_{j\ell} - Y'_{j\ell}),$$

where $X_{j\ell} \sim$ Bernoulli(q) and $Y_{j\ell}, Y'_{j\ell} \sim$ Bernoulli$(1-\alpha)$, all of which are independent. Thus, we have $\tilde{\mathbb{E}}[A] = 0$ and

$$\tilde{\mathbb{E}}[A^2] = \sum_{j\in I} \sum_{\ell\in[n]\setminus\mathcal{B}_{G_0}(i,m+1)} \mathbb{E}\Big[X_{j\ell}^2(Y_{j\ell} - Y'_{j\ell})^2\Big] = 2q\alpha(1-\alpha)\tilde{n}|I| = 2\alpha p\,\tilde{n}|I|.$$

Moreover, the random variables $X_{j\ell}(Y_{j\ell} - Y'_{j\ell})$ take values in $\{-1, 0, 1\}$ and are i.i.d. with mean zero and variance $2\alpha p$. Hence Bernstein's inequality yields the desired tail bound for A. □

Since i is fixed, we drop the argument (i) in $T_s(i)$, $f(i)_s$, $R_s(i)$, etc. to ease the notation below when there is no ambiguity. Recall that R_s and R'_s are defined by (22); moreover, η_s and η'_s are defined by (25). Let f and v be the signature vector and the variance vector respectively given by `VertexSignature`(G, i, m) (Algorithm 1), and let f' and v' be given by `VertexSignature`(G', i, m).

Lemma 5.9 (Entrywise difference between signatures). *Condition further on a realization of edges between $R_s \cup R'_s$ and $\mathcal{S}_{G_0}(i, m+2)$ in the graphs G_0, G, and G'. Let $\hat{\mathbb{P}}$ and $\hat{\mathbb{E}}$ denote the conditional probability and expectation respectively. Then, for $s \in J$, we have*

$$f_s - f'_s = Z_s + \Delta_s$$

for a random variable Z_s and a deterministic quantity Δ_s satisfying

- $\hat{\mathbb{E}}[Z_s] = 0$;
- $\hat{\mathbb{E}}[Z_s^2] \le \mathsf{v}_s + \mathsf{v}'_s - 2\tilde{n}\,p(1-p-\alpha)\big|\mathcal{N}_G(T_s) \cap \mathcal{N}_{G'}(T'_s) \cap \mathcal{S}_{G_0}(i, m+1)\big|$;
- $\hat{\mathbb{P}}\big\{|Z_s| \ge t\big\} \le 2\exp\Big(\frac{-t^2/2}{\hat{E}[Z_s^2]+t/3}\Big)$;
- $|\Delta_s| \le |\eta_s| + |\eta'_s| + 2n^{0.2}p$.

Moreover, the random variables Z_s are conditionally independent for different $s \in J$.

Proof First, note that because of the further conditioning on the edges between $R_s \cup R'_s$ and $\mathcal{S}_{G_0}(i, m+2)$, the quantities η_s and η'_s defined in (25) become deterministic.

As $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree by Condition (B1), for every $v \in \mathcal{S}_G(i, m)$, we have

$$\mathcal{N}_G(v) \cap \mathcal{S}_G(i, m+1) = \mathcal{N}_G(v) \cap \mathcal{S}_{G_0}(i, m+1).$$

Therefore, it holds that

$$\begin{aligned} f_s &= \sum_{j\in\mathcal{N}_G(T_s)\cap\mathcal{S}_G(i,m+1)} \big(\deg_G(j) - 1 - np\big) \\ &= \sum_{j\in\mathcal{N}_G(T_s)\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_G(j) - 1 - \tilde{n}p\big) + (\tilde{n}-n)p\,\big|\mathcal{N}_G(T_s)\cap\mathcal{S}_{G_0}(i,m+1)\big|. \end{aligned}$$

Furthermore, in view of the partition

$$\mathcal{N}_G(T_s) = \big(\mathcal{N}_G(T_s)\cap\mathcal{N}_{G'}(T'_s)\big) \cup \big(\mathcal{N}_G(T_s)\backslash\mathcal{N}_{G'}(T'_J)\big) \cup \big(\mathcal{N}_G(T_s)\cap\mathcal{N}_{G'}(T'_{J\backslash\{s\}})\big),$$

we obtain

$$\begin{aligned} f_s = &\sum_{j\in\mathcal{N}_G(T_s)\cap\mathcal{N}_{G'}(T'_s)\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_G(j) - 1 - \tilde{n}p\big) \\ &+ \sum_{j\in(\mathcal{N}_G(T_s)\backslash\mathcal{N}_{G'}(T'_J))\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_G(j) - 1 - \tilde{n}p\big) \\ &+ \eta_s + (\tilde{n}-n)p\,\big|\mathcal{N}_G(T_s)\cap\mathcal{S}_{G_0}(i,m+1)\big|. \end{aligned}$$

Consequently,

$$f_s - f'_s = Z_s + \Delta_s,$$

where

$$\begin{aligned} Z_s := &\sum_{j\in\mathcal{N}_G(T_s)\cap\mathcal{N}_{G'}(T'_s)\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_G(j) - \deg_{G'}(j)\big) \\ &+ \sum_{j\in(\mathcal{N}_G(T_s)\backslash\mathcal{N}_{G'}(T'_J))\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_G(j) - 1 - \tilde{n}p\big) \\ &- \sum_{j\in(\mathcal{N}_{G'}(T'_s)\backslash\mathcal{N}_G(T_J))\cap\mathcal{S}_{G_0}(i,m+1)} \big(\deg_{G'}(j) - 1 - \tilde{n}p\big) \end{aligned}$$

and

$$\Delta_s := \eta_s - \eta'_s + (\tilde{n}-n)p\,\Big(\big|\mathcal{N}_G(T_s)\cap\mathcal{S}_{G_0}(i,m+1)\big| - \big|\mathcal{N}_{G'}(T'_s)\cap\mathcal{S}_{G_0}(i,m+1)\big|\Big).$$

For the deterministic quantity Δ_s, we have

$$|\Delta_s| \le |\eta_s| + |\eta'_s| + n^{0.1}p\cdot 2n^{0.1} = |\eta_s| + |\eta'_s| + 2n^{0.2}p.$$

For the random variable Z_s, it is not hard to see that the three sums in its definition are over disjoint sets, so they are independent. Moreover, for Z_s, the probabilities $\tilde{\mathbb{P}}$ and $\hat{\mathbb{P}}$ coincide, because the extra conditioning for $\hat{\mathbb{P}}$ is on edges disjoint from those

used to define Z_s. The means, variances, and tail bounds for the three terms of Z_s can be derived using Lemma 5.8. Namely, we obtain $\hat{\mathbb{E}}[Z_s] = 0$ and

$$\begin{aligned}
\hat{\mathbb{E}}[Z_s^2] &= 2\alpha p\tilde{n}\big|\mathcal{N}_G(T_s) \cap \mathcal{N}_{G'}(T_s') \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\quad + p(1-p)\tilde{n}\big|\big(\mathcal{N}_G(T_s)\backslash\mathcal{N}_{G'}(T_J')\big) \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\quad + p(1-p)\tilde{n}\big|\big(\mathcal{N}_{G'}(T_s')\backslash\mathcal{N}_G(T_J)\big) \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\le p(1-p)\tilde{n}\Big(\big|\mathcal{N}_G(T_s) \cap \mathcal{S}_{G_0}(i, m+1)\big| + \big|\mathcal{N}_{G'}(T_s') \cap \mathcal{S}_{G_0}(i, m+1)\big|\Big) \\
&\quad - \big(2p(1-p)\tilde{n} - 2\alpha p\tilde{n}\big)\big|\mathcal{N}_G(T_s) \cap \mathcal{N}_{G'}(T_s') \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\le \mathsf{v}_s + \mathsf{v}_s' - 2\tilde{n}p(1-p-\alpha)\big|\mathcal{N}_G(T_s) \cap \mathcal{N}_{G'}(T_s') \cap \mathcal{S}_{G_0}(i, m+1)\big|,
\end{aligned}$$

where the first inequality holds because all the neighborhoods in the three terms are disjoint and contained in either $\mathcal{N}_G(T_s)$ or $\mathcal{N}_{G'}(T_s')$. The tail bound for Z_s follows again from Bernstein's inequality as in Lemma 5.8. The conditional independence of Z_s for different $s \in J$ follows from the disjointness of sets that we sum over in the definition of Z_s. □

Lemma 5.10 (Hoeffding's inequality with truncation). *Let $X_1, \ldots, X_N$ be independent random variables satisfying that $\big|\mathbb{E}[X_i]\big| \le \tau$ for $\tau > 0$ and that*

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

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 $\delta \in (0, 0.1)$,

$$\begin{aligned}
\mathbb{P}\Big\{\Big|\sum_{i=1}^N \big(X_i^2 - \mathbb{E}[X_i^2]\big)\Big| \ge C\log(N/\delta)\sqrt{N\log(1/\delta)} \\
+ C\tau\Big(\sqrt{N\log(1/\delta)} + \log(1/\delta)\Big)\Big\} \le \delta.
\end{aligned}$$

Proof It is easily seen that

$$\begin{aligned}
X_i^2 - \mathbb{E}[X_i^2] &= \big(X_i - \mathbb{E}[X_i]\big)^2 - \mathbb{E}\big[\big(X_i - \mathbb{E}[X_i]\big)^2\big] + 2\,\mathbb{E}[X_i]\big(X_i - \mathbb{E}[X_i]\big) \\
&= \big(Y_i^2 - \mathbb{E}[Y_i^2]\big) + 2\,\mathbb{E}[X_i]\,Y_i
\end{aligned}$$

where $Y_i := X_i - \mathbb{E}[X_i]$. It suffices to control the sum of the above two terms over $i \in [N]$.

Let us first control $\sum_{i=1}^N \big(Y_i^2 - \mathbb{E}[Y_i^2]\big)$. For $M > 0$, we have

$$\begin{aligned}
\mathbb{E}[Y_i^2] - \mathbb{E}[Y_i^2 \cdot \mathbb{1}\{|Y_i| \le M\}] &= \mathbb{E}[Y_i^2 \cdot \mathbb{1}\{|Y_i| > M\}] \\
&= \int_M^\infty 2t \cdot \mathbb{P}\{|Y_i| > t\}\,dt
\end{aligned}$$

$$= 4\int_M^\infty t\exp\Big(\frac{-ct^2}{1+t}\Big)\,dt$$
$$\le 16\frac{cM+1}{c^2}\exp(-cM/2) \le (\delta/N)^{100},$$

if $\delta \in (0, 0.1)$ and $M = C_1\log(N/\delta)$ for a sufficiently large constant $C_1 = C_1(c) > 0$. Moreover, by a union bound, $Y_i^2 = Y_i^2\cdot\mathbb{1}\{|Y_i| \le M\}$ for all $i \in [N]$ with probability at least $1-(\delta/N)^{100}$ if C_1 is sufficiently large. Thus, by Hoeffding's inequality applied to $Y_i^2\cdot\mathbb{1}\{|Y_i| \le M\}$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^N \big(Y_i^2 - \mathbb{E}[Y_i^2\cdot\mathbb{1}\{|Y_i| \le M\}]\big)\Big| \ge 2M\sqrt{N\log(1/\delta)}\Big\} \le \delta.$$

Combining the above two displays yields

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^N \big(Y_i^2 - \mathbb{E}[Y_i^2]\big)\Big| \ge 3C_1\log(N/\delta)\sqrt{N\log(1/\delta)}\Big\} \le \delta.$$

Next, we turn to the term $\sum_{i=1}^N \mathbb{E}[X_i]\,Y_i$. Since $\big|\mathbb{E}[X_i]\big| \le \tau$, the variable $\mathbb{E}[X_i]\,Y_i$ is sub-exponential with parameter $C_2\tau$ for a universal constant $C_2 > 0$. Bernstein's inequality then implies that

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^N \mathbb{E}[X_i]\,Y_i\Big| \ge C_3\tau\Big(\sqrt{N\log(1/\delta)} + \log(1/\delta)\Big)\Big\} \le \delta$$

for a universal constant $C_3 > 0$.

The above two parts combined complete the proof. □

Lemma 5.11 (Difference between signatures). *For any constants $C, D, K, K' > 0$, there exist constants $n_0, R, \kappa > 0$ with the following property. Suppose that*

$$n \ge n_0, \qquad \log n \le nq \le n^{\frac{1}{R\log\log n}}, \qquad \alpha \in (0,\kappa),$$
$$\log\big(w^4(\log n)^2\big) \le m \le C\log\log n, \qquad w \ge (\log n)^4.$$

Moreover, suppose that Conditions (B1) to (B7) hold with constants $K, \kappa > 0$ and a subset J with $|J| = 2w$. Consider the same conditioning as in Lemma 5.9, such that η_s and η_s' satisfy (24) for a constant $K' > 0$. Then it holds with conditional probability at least $1-n^{-D}$ that

$$\sum_{s\in J}\frac{\big(f_s - f_s'\big)^2}{\mathsf{v}_s + \mathsf{v}_s'} \le 2w\big(1-(1-9\kappa)^m\big).$$

Proof Let

$$I := \Big\{ j \in \mathcal{S}_{G_0}(i,m) : \deg_{G\cap G'}(j) \le \frac{2}{3}np(1-\alpha) \Big\}.$$

By Conditions (B6) and (B7), we have that for any $s \in \{-1,1\}^m$,

$$|I| \le (nq/3)^m \le \frac{1}{2}(np/2)^m(1-8\kappa)^m \le \frac{1}{2}|T_s \cap T_s'|$$

if $\kappa, \alpha \in (0, 0.01)$ and $m \ge 3$. Since $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree, it is easy to see that

$$\begin{aligned} &\big|\mathcal{N}_G(T_s) \cap \mathcal{N}_{G'}(T_s') \cap \mathcal{S}_{G_0}(i,m+1)\big| \\ &\quad \ge \sum_{j\in(T_s\cap T_s')\setminus I} \big(\deg_{G\cap G'}(j) - 1\big) \ge \frac{1}{2}np \cdot \frac{1}{2}(np/2)^m(1-8\kappa)^m \\ &\quad = \frac{1}{2}(np/2)^{m+1}(1-8\kappa)^m. \end{aligned} \tag{26}$$

Similarly, we also have

$$\begin{aligned} \mathsf{v}_s &= np(1-p)\big|\mathcal{N}_G(T_s) \cap \mathcal{S}_G(i,m+1)\big| \ge np(1-p) \sum_{j\in T_s\setminus I} \big(\deg_G(j) - 1\big) \\ &\ge np(1-p) \cdot \frac{1}{2}(np/2)^m(1-8\kappa)^m \cdot \frac{1}{2}np = (1-p)(np/2)^{m+2}(1-8\kappa)^m. \end{aligned}$$

On the other hand, by Conditions (B3) and (B4),

$$\mathsf{v}_s = np(1-p)\big|\mathcal{N}_G(T_s) \cap \mathcal{S}_G(i,m+1)\big| \le np \sum_{j\in T_s} \deg_G(j) \le K^2 \frac{(np)^{m+2}}{2^m}.$$

The same estimates also hold for v_s'.

We will apply Lemma 5.10 with

$$X_s := \frac{f_s - f_s'}{\sqrt{\mathsf{v}_s + \mathsf{v}_s'}} = \frac{Z_s + \Delta_s}{\sqrt{\mathsf{v}_s + \mathsf{v}_s'}},$$

where Z_s and Δ_s satisfy the conclusion of Lemma 5.9. Towards that end, let us first establish some estimates for the mean and variance of X_s. Using Lemma 5.9, (24), and the above estimates for v_s and v_s', we obtain

$$\begin{aligned} \big|\hat{\mathbb{E}}[X_s]\big| &= \frac{|\Delta_s|}{\sqrt{\mathsf{v}_s + \mathsf{v}_s'}} \\ &\le \frac{|\eta_s| + |\eta_s'| + 2n^{0.2}p}{\sqrt{2(1-p)(np/2)^{m+2}(1-8\kappa)^m}} \end{aligned}$$

$$\le \frac{2K'\frac{(np)^{m/2+1}}{2^m}w^2\sqrt{\log n}+2n^{0.2}p}{\sqrt{2(1-p)(np/2)^{m+2}(1-8\kappa)^m}} \le \frac{5K'w^2\sqrt{\log n}}{2^{m/2}(1-8\kappa)^{m/2}} \quad \text{for } s\in\tilde{J}(i),$$

$$\big|\hat{\mathbb{E}}[X_s]\big| = \frac{|\Delta_s|}{\sqrt{\mathsf{v}_s+\mathsf{v}'_s}} \le \frac{2K'\frac{(np)^{m/2+1}}{2^{m/2}}\sqrt{\log n}+2n^{0.2}p}{\sqrt{2(1-p)(np/2)^{m+2}(1-8\kappa)^m}} \le \frac{5K'\sqrt{\log n}}{(1-8\kappa)^{m/2}} \quad \text{for } s\in J, \tag{27}$$

and

$$\begin{aligned}\mathrm{Var}(X_s) = \frac{\hat{\mathbb{E}}[Z_s^2]}{\mathsf{v}_s+\mathsf{v}'_s} &\le \frac{\mathsf{v}_s+\mathsf{v}'_s-2\tilde{n}p(1-p-\alpha)\big|\mathcal{N}_G(T_s)\cap\mathcal{N}_{G'}(T'_s)\cap\mathcal{S}_{G_0}(i,m+1)\big|}{\mathsf{v}_s+\mathsf{v}'_s}\\ &\le 1-\frac{\tilde{n}p(1-p-\alpha)(np/2)^{m+1}(1-8\kappa)^m}{K^2(np)^{m+2}/2^m} \le 1-\frac{(1-8\kappa)^m}{3K^2}\end{aligned}$$

by Lemma 5.9 and (26) if $\alpha \le 0.1$. Therefore,

$$\begin{aligned}\hat{\mathbb{E}}[X_s^2] = \frac{\hat{\mathbb{E}}[Z_s^2]+\Delta_s^2}{\mathsf{v}_s+\mathsf{v}'_s} &\le 1-\frac{(1-8\kappa)^m}{3K^2}+\frac{25(K')^2w^4\log n}{2^m(1-8\kappa)^m}\\ &\le 1-\frac{(1-8\kappa)^m}{4K^2} \quad \text{for } s\in\tilde{J}(i)\end{aligned}$$

if $\log\big(w^4(\log n)^2\big) \le m \le C\log\log n$, $\kappa > 0$ is sufficiently small depending on C, and $n \ge n_0 = n_0(K, K', \kappa)$, and

$$\hat{\mathbb{E}}[X_s^2] = \frac{\hat{\mathbb{E}}[Z_s^2]+\Delta_s^2}{\mathsf{v}_s+\mathsf{v}'_s} \le 1-\frac{(1-8\kappa)^m}{3K^2}+\frac{25(K')^2\log n}{(1-8\kappa)^m} \le \frac{26(K')^2\log n}{(1-8\kappa)^m} \quad \text{for } s\in J.$$

By the bound $\hat{\mathbb{P}}\big\{|Z_s|\ge t\big\} \le 2\exp\left(\frac{-t^2/2}{\hat{E}[Z_s^2]+t/3}\right)$ in Lemma 5.9 and that $\mathrm{Var}(X_s) \le 1$, we see that $\hat{\mathbb{P}}\big\{|X_s-\hat{\mathbb{E}}[X_s]|\ge t\big\} \le 2\exp\left(\frac{-t^2/2}{1+t/3}\right)$. Therefore, Lemma 5.10 can be applied to show that, with conditional probability at least $1-n^{-D}$,

$$\sum_{s\in J}X_s^2 \le \sum_{s\in J}\hat{\mathbb{E}}[X_s^2]+C_1\sqrt{w}(\log n)^{3/2}+C_1\Big(\max_{s\in J}\big|\hat{\mathbb{E}}[X_s]\big|\Big)\big(\sqrt{w\log n}+\log n\big)$$

for a constant $C_1 > 0$ depending on D, where we recall $|J| = 2w$. Moreover, by Condition (B5) and the above bounds on $\hat{\mathbb{E}}[X_s^2]$, we have

$$\begin{aligned}\sum_{s\in J}\hat{\mathbb{E}}[X_s^2] &= \sum_{s\in\tilde{J}(i)}\hat{\mathbb{E}}[X_s^2]+\sum_{s\in J\setminus\tilde{J}(i)}\hat{\mathbb{E}}[X_s^2]\\ &\le 2w\Big(1-\frac{(1-8\kappa)^m}{4K^2}\Big)+(\log n)^2\frac{26(K')^2\log n}{(1-8\kappa)^m} \le 2w\Big(1-\frac{(1-8\kappa)^m}{5K^2}\Big)\end{aligned}$$

if $m \le C \log\log n$, $\kappa > 0$ is sufficiently small depending on C, $w \ge (\log n)^4$, and $n \ge n_0$. The above two bounds together with (27) imply that

$$\begin{aligned}\sum_{s\in J} X_s^2 &\le 2w\Big(1 - \frac{(1-8\kappa)^m}{5K^2}\Big) + C_1\sqrt{w}(\log n)^{3/2} \\ &\quad + C_1\frac{5K'\sqrt{\log n}}{(1-8\kappa)^{m/2}}\big(\sqrt{w\log n} + \log n\big) \\ &\le 2w\Big(1 - \frac{(1-8\kappa)^m}{6K^2}\Big)\end{aligned}$$

if, again, $m \le C\log\log n$, $\kappa > 0$ is sufficiently small depending on C, $w \ge (\log n)^4$, and $n \ge n_0$. Finally, if $m \ge \log\log n$ and $n \ge n_0 = n_0(K,\kappa)$, then

$$\frac{(1-8\kappa)^m}{6K^2} \ge \frac{(1-8\kappa)^m}{(1+\kappa)^m} \ge (1-9\kappa)^m,$$

so the proof is complete. □

5.4 Conclusion

We summarize the result of this section in the following proposition.

Proposition 5.12 (Difference between signatures of typical correct pairs). *For any constants C, $D > 0$, there exist constants n_0, R, α_0, $c > 0$ with the following property. Let J be a uniform random subset of $\{-1,1\}^s$ of cardinality $2w$ for $w \in \mathbb{N}$. Suppose that*

$$\begin{aligned} &n \ge n_0, \qquad \log n \le nq \le n^{\frac{1}{R\log\log n}}, \qquad \alpha \in (0,\alpha_0), \\ &\log\big(w^4(\log n)^2\big) \le m \le C\log\log n, \qquad w \ge (\log n)^4. \end{aligned}$$

Then with probability at least $1 - n^{-D}$, for at least $n - n^{1-c}$ vertices $i \in [n]$, we have

$$\sum_{s\in J} \frac{\big(f_s(i) - f_s'(i)\big)^2}{v_s(i) + v_s'(i)} \le 2w\Big(1 - \frac{1}{(\log n)^{0.1}}\Big). \tag{28}$$

Proof We first check that, with probability at least $1 - n^{-D-1}$, Conditions (B1) to (B7) hold for at least $n - n^{1-c}$ vertices $i \in [n]$ with constants $\kappa > 0$ to be chosen and $K = K(C,D) > 0$, where $c > 0$ depends on C, D, and κ:

1. Condition (B1) is the same as Condition (A1), so the result follows from Proposition 4.12.
2. For Condition (B2), we apply Lemma 4.2 and the relations $n \ge n_0$, $\log n \le nq \le n^{\frac{1}{R\log\log n}}$, and $m \le C\log\log n$, where n_0 and R depend on C and D, to obtain the following: With probability at least $1 - n^{-D-2}$, it holds for all $i \in [n]$ that $|\mathcal{B}_{G_0}(i, m+1)| \le K(np)^{m+1} \le n^{0.1}$ where K depends on D.

3. Condition (B3) is standard and in fact holds for all vertices $j \in [n]$ with probability at least $1 - n^{-D-2}$ for a constant $K > 0$ depending only on D.
4. By Lemma 4.3, with probability at least $1 - n^{-D-2}$, Condition (B4) holds for all vertices i such that $G_0(\mathcal{B}_{G_0}(i, m+1))$ is a tree, that is, whenever Condition (B1) holds.
5. Condition (B5) is the conclusion of Lemma 5.6, so it suffices to check the three assumptions in that lemma, but those are already guaranteed by Conditions (B1), (B3), and (B4) respectively. Therefore, Condition (B5) holds up to a possible change of the constant K.
6. Condition (B6) is very similar to Condition (A4), with $l = m$, $\kappa = 1/3$, and the graph G_0 replaced by $G \cap G'$ which is a $G(n, p(1-\alpha))$ random graph. As a result, a straightforward modification of Point 4 in the proof of Proposition 4.12 yields result. Namely, we obtain the bound $\left|\left\{j \in \mathcal{S}_{G_0}(i,m) : \deg_{G\cap G'}(j) \le \frac{2}{3}np(1-\alpha)\right\}\right| \le \frac{1}{K\cdot 3^m}|\mathcal{S}_{G_0}(i,m)|$ for at least $n - n^{1-2c}$ vertices $i \in [n]$ with probability at least $1 - n^{-D-2}$, but $|\mathcal{S}_{G_0}(i,m)| \le K(nq)^m$ by Lemma 4.2.
7. Given the constants κ, C, and D, we can choose δ and K according to Proposition 4.12, and then choose α_0 according to Proposition 5.1 to obtain Condition (B7) for at least $n - n^{1-2c}$ typical vertices $i \in [n]$ with probability at least $1 - n^{-D-2}$.

Next, for any $i \in [n]$ satisfying Conditions (B1) to (B7), we can apply Lemma 5.7 to obtain that $\eta_s(i)$ and $\eta'_s(i)$ satisfy (24) with probability at least $1 - n^{-D-2}$ for a constant $K' = K'(D, K) > 0$. Then by Lemma 5.11, we obtain (28) for any such vertex i with probability at least $1 - n^{-D-2}$, if $\kappa > 0$ is chosen according to the lemma and is sufficiently small depending on C so that $(1-9\kappa)^m \ge \frac{1}{(\log n)^{0.1}}$. □

6 Signatures of wrong pairs of vertices

The structure of proofs in this subsection is similar to that in the previous subsection. However, the technical details are slightly more involved when considering two (possibly intersecting) neighborhoods $\mathcal{N}_{G_0}(i, m+1)$ and $\mathcal{N}_{G_0}(i', m+1)$, where i and i' are distinct vertices.

6.1 Sparsification

The following lemma is in the same spirit as Lemma 5.6. In particular, note the extra factor $\left(\frac{w}{np(1-\alpha)^m} + \sqrt{w \log n}\right)$ in (29b) compared to (29a) (a trivial bound would give a factor $|J| = 2w$ instead). This will be crucial to controlling $\zeta_s(i, i')$ and $\zeta'_s(i, i')$ in Lemma 6.2 and subsequent estimates in Lemma 6.4.

Lemma 6.1 (Sparsification for a pair of vertices). *For constants $C_1, C_2, C_3, D > 0$, there exists $K = K(C_1, C_2, C_3, D) > 0$ with the following property. Let J be a uniform random subset of $\{-1, 1\}^m$ of cardinality $2w \ge \log n$ for $w \in \mathbb{N}$. With respect to the randomness of J, the following holds with probability at least $1 - n^{-D}$ for any distinct vertices $i, i' \in [n]$: If*

- *$G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ and $G_0\big(\mathcal{B}_{G_0}(i', m+1)\big)$ are trees,*

- $\deg_G(j) \vee \deg_{G'}(j) \le C_1 np$ *for all* $j \in \mathcal{B}_{G_0}(i,m) \cup \mathcal{B}_{G_0}(i,m)$,
- $|T_s(i)| \vee |T'_s(i')| \le C_2\big(\frac{np}{2}\big)^m$ *for all* $s \in \{-1,1\}^m$, *and*
- $|\mathcal{B}_{G_0}(i,m+1) \cap \mathcal{B}_{G_0}(i',m+1)| \le C_3(nq)^m$,

then

$$\Big(\max_{s\in J}|L_s(i,i')|\Big) \vee \Big(\max_{s\in J}|L'_s(i,i')|\Big) \le K\frac{(np)^{m+1}}{2^m}, \tag{29a}$$

$$\Big(\sum_{s\in J}|L_s(i,i')|\Big) \vee \Big(\sum_{s\in J}|L'_s(i,i')|\Big) \le K\frac{(np)^{m+1}}{2^m}\Big(\frac{w}{np(1-\alpha)^m} + \sqrt{w\log n}\Big), \tag{29b}$$

where

$$L_s(i,i') := \mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{B}_{G_0}(i',m+1) \cap \mathcal{S}_{G_0}(i,m+1), \tag{30a}$$

$$L'_s(i,i') := \mathcal{N}_{G'}\big(T'_s(i')\big) \cap \mathcal{B}_{G_0}(i,m+1) \cap \mathcal{S}_{G_0}(i',m+1). \tag{30b}$$

Proof Fix distinct vertices $i, i' \in [n]$ such that the four conditions in the lemma hold. For $s \in \{-1,1\}^m$, let

$$a_s := \big|L_s(i,i')\big|.$$

By the conditions $|T_s(i)| \vee |T'_s(i')| \le C_2\big(\frac{np}{2}\big)^m$ for all $s \in \{-1,1\}^m$ and $\deg_G(j) \vee \deg_{G'}(j) \le C_1 np$ for all $j \in \mathcal{B}_{G_0}(i,m) \cup \mathcal{B}_{G_0}(i,m)$, we have

$$a_s \le \sum_{j\in T_s(i)} \deg_G(j) \le C_1C_2\frac{(np)^{m+1}}{2^m}.$$

The same bound also holds for $L'_s(i,i')$. Hence (29a) is proved.

In addition, we have

$$\sum_{s\in\{-1,1\}^m} a_s \le \big|\mathcal{B}_{G_0}(i,m+1) \cap \mathcal{B}_{G_0}(i',m+1)\big| \le C_3(nq)^m.$$

Since J is a uniform random subset of $\{-1,1\}^m$, Bernstein's inequality for sampling without replacement implies that, with probability at least $1 - n^{-D-3}$,

$$\begin{aligned}\sum_{s\in J} a_s &\le \frac{|J|}{2^m}\sum_{s\in\{-1,1\}^m} a_s + K_1\sqrt{\frac{|J|}{2^m}\sum_{s\in\{-1,1\}^m} a_s^2\cdot\log n} + K_1\max_{s\in\{-1,1\}^m}|a_s|\cdot\log n \\ &\le K_2\frac{|J|}{2^m}(nq)^m + K_2\frac{(np)^{m+1}}{2^m}\sqrt{|J|\cdot\log n} + K_2\frac{(np)^{m+1}}{2^m}\log n \\ &\le 3K_2\frac{(np)^{m+1}}{2^m}\Big(\frac{w}{np(1-\alpha)^m} + \sqrt{w\log n}\Big)\end{aligned}$$

for $K_1, K_2 > 0$ depending on C_1, C_2, C_3, and D, where we used that $|J| = 2w \geq \log n$. The same bound also holds for $\sum_{s\in J} |L'_s(i,i')|$, so (29b) is proved.

A union bound over all distinct $i, i' \in [n]$ then completes the proof. □

Lemma 6.2 (Small overlaps). *For any $C, D > 1$, there exists $c \in (0, 1/2)$ and $K, K', R, n_0 > 1$ depending on C and D with the following property. Let J be a uniform random subset of $\{-1, 1\}^m$ of cardinality $2w \geq \log n$ for $w \in \mathbb{N}$. If*

$$n \geq n_0, \qquad \log n \leq nq \leq n^{\frac{1}{R \log\log n}}, \qquad m \leq C \log\log n,$$

then with probability at least $1 - n^{-D}$, there is a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that the following holds. For all distinct $i, i' \in \mathcal{I}$, (29a) and (29b) hold, and

$$\Big(\max_{s\in J} \big|\zeta_s(i,i')\big|\Big) \vee \Big(\max_{s\in J} \big|\zeta'_s(i,i')\big|\Big) \leq K' \frac{(np)^{m/2+1}}{2^{m/2}} \sqrt{\log n}, \tag{31}$$

where

$$\zeta_s(i,i') := \sum_{j\in L_s(i,i')} \big(\deg_G(j) - 1 - np\big)$$
$$\text{and} \quad \zeta'_s(i,i') := \sum_{j\in L'_s(i,i')} \big(\deg_{G'}(j) - 1 - np\big). \tag{32}$$

Proof Since we will apply Lemma 6.1, let us first show that, with probability at least $1 - n^{-D-1}$, there is a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that the four conditions in Lemma 6.1 hold for any distinct vertices $i, i' \in \mathcal{I}$:

- Proposition 4.12 shows that with probability at least $1 - n^{-D}$, at least $n - n^{1-c}$ vertices satisfy Condition (A1). The same proof with a slight change of constants implies that, with probability at least $1 - n^{-D-2}$, there is a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that $G_0\big(\mathcal{B}_{G_0}(i, 3m+3)\big)$ is a tree for all $i \in \mathcal{I}$. (Here the radius of the neighborhood is chosen to be $3m + 3$ which is larger than $m + 1$ required in the first condition of Lemma 6.1; this is because $3m + 3$ is needed in the fourth condition below.)
- As before, we have $\deg_G(j) \vee \deg_{G'}(j) \leq C_1 np$ for all $j \in [n]$ with probability at least $1 - n^{-D-2}$ where $C_1 > 0$ depends on D.
- By Lemma 4.3, with probability at least $1 - n^{-D-2}$, we have $|T_s(i)| \vee |T'_s(i')| \leq C_2\big(\frac{np}{2}\big)^m$ for all $s \in \{-1, 1\}^m$ and all vertices i such that $G_0(\mathcal{B}_{G_0}(i, m+1))$ is a tree, where $C_2 > 0$ depends on C and D.
- By Lemma 4.2, it holds with probability at least $1 - n^{-D-2}$ that, for any distinct vertices $i, i' \in [n]$ such that $G_0\big(\mathcal{B}_{G_0}(i, 3m+3)\big)$ is tree, we have $|\mathcal{B}_{G_0}(i, m+1) \cap \mathcal{B}_{G_0}(i', m+1)| \leq C_3(nq)^m$ where $C_3 > 0$ depends on D.

Therefore, by Lemma 6.1, we obtain (29a) and (29b) for all distinct $i, i' \in \mathcal{I}$ with probability at least $1 - 2n^{-D-1}$, where $|\mathcal{I}| \geq n - n^{1-c}$.

The rest of the proof is split into three cases according to the value of $d := \mathrm{dist}_{G_0}(i, i')$. We focus on proving (31) for $\zeta_s(i, i')$, as the same argument also works for $\zeta_s'(i, i')$.

Case 1: $d \le m+1$. In this case, we have $i' \in \mathcal{B}_{G_0}(i, m+1)$. Let us condition on an instance of $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$, $G\big(\mathcal{B}_{G_0}(i, m+1)\big)$, and $G'\big(\mathcal{B}_{G_0}(i, m+1)\big)$ such that $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree. Then $L_s(i, i')$ defined in (30) is equal to

$$L_s(i, i') = \mathcal{N}_G\big(T_s(i)\big) \cap \mathcal{S}_{G_0}(i', m+1-d) \cap \mathcal{S}_{G_0}(i, m+1).$$

Under this conditioning, the random variables $\deg_G(j) - 1$, $j \in L_s(i, i')$, are conditionally independent Binomial$(\tilde{n}, p)$ where

$$\tilde{n} := n - \big|\mathcal{B}_{G_0}(i, m+1)\big| \ge n - n^{0.1}$$

as before. Consequently, $\sum_{j \in L_s(i,i')} \big(\deg_G(j) - 1\big) \sim \text{Binomial}(\tilde{n}|L_s(i, i')|, p)$, and so

$$\begin{aligned} |\zeta_s(i, i')| &= \Big| \sum_{j \in L_s(i,i')} \big(\deg_G(j) - 1 - np\big) \Big| \\ &\le K_2\big(\sqrt{np|L_s(i, i')| \log n} + \log n\big) + (n - \tilde{n})p\,|L_s(i, i')| \\ &\le K_3\sqrt{np|L_s(i, i')| \log n} \end{aligned}$$

with conditional probability at least $1 - n^{-D-2}$ for constants $K_2, K_3 > 0$ depending on D. By a union bound over $s \in \{-1, 1\}^m$ and $i, i' \in [n]$ together with (29a), we obtain (31) for $\zeta_s(i, i')$.

Case 2: $m+1 < d \le 2m+2$. In this case, we have $i' \notin \mathcal{B}_{G_0}(i, m+1)$. Let us condition on an instance of $G_0\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$, $G\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$, and $G'\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$ such that $G_0\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$ is a tree. Then there is a unique path γ in $G_0\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$ connecting i to i', and γ passes through a unique vertex $v \in \mathcal{S}_{G_0}(i, m+1)$. Note that v is adjacent to exactly two vertices in $\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)$, because otherwise there would be a cycle in $G_0\big(\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big)$. For the same reason, if $j \in \mathcal{S}_{G_0}(i, m+1) \cap \mathcal{B}_{G_0}(i', m+1)$ and $j \ne v$, then j is adjacent to exactly one vertex in $\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)$.

In view of this structure, we have the following observation under the above conditioning: If $v \notin L_s(i, i')$, then the random variables

$$\deg_G(j) - 1, \;\; j \in L_s(i, i'),$$

are conditionally independent Binomial$(\bar{n}, p)$, where

$$\bar{n} := n - \big|\mathcal{B}_{G_0}(i, m+1) \cup \mathcal{B}_{G_0}(i', m+1)\big| \ge n - n^{0.1}. \tag{33}$$

On the other hand, if $v \in L_s(i, i')$, then

$$\deg_G(v) - 2 \quad \text{and} \quad \deg_G(j) - 1, \ \ j \in L_s(i, i') \backslash \{v\},$$

are conditionally independent Binomial$(\bar{n}, p)$.

Consequently, $\sum_{j \in L_s(i,i')} \big(\deg_G(j) - 1\big) - \mathbb{1}\{v \in L_s(i, i')\}$ is Binomial$(\bar{n}|L_s(i, i')|, p)$, and so

$$\begin{aligned} |\zeta_s(i, i')| &\le \Big| \sum_{j \in L_s(i,i')} \big(\deg_G(j) - 1\big) - \mathbb{1}\{v \in L_s(i, i')\} \Big| + 1 \\ &\le K_2\big(\sqrt{np|L_s(i, i')| \log n} + \log n\big) + (n - \bar{n})p\,|L_s(i, i')| + 1 \\ &\le K_3\sqrt{np|L_s(i, i')| \log n} \end{aligned}$$

with conditional probability at least $1 - n^{-D-2}$ for constants $K_2, K_3 > 0$ depending on D. By a union bound over $s \in \{-1, 1\}^m$ and $i, i' \in [n]$ together with (29a), we obtain (31) for $\zeta_s(i, i')$.

Case 3: $d > 2m + 2$. This case is trivial because $L_s(i, i') \subset \mathcal{B}_{G_0}(i, m + 1) \cap \mathcal{B}_{G_0}(i', m + 1) = \varnothing$ so that $\zeta_s(i, i') = 0$. □

6.2 Difference between signatures of a typical pair

Throughout this subsection, we fix distinct vertices $i, i' \in [n]$ and a subset $J \subset \{-1, 1\}^m$ of cardinality $2w$ for $w \in \mathbb{N}$. Moreover, let us condition on an instance of the three subgraphs

$$G_0\big(\mathcal{B}_{G_0}(i, m + 1) \cup \mathcal{B}_{G_0}(i', m + 1)\big), \ G\big(\mathcal{B}_{G_0}(i, m + 1) \cup \mathcal{B}_{G_0}(i', m + 1)\big),$$
$$G'\big(\mathcal{B}_{G_0}(i, m + 1) \cup \mathcal{B}_{G_0}(i', m + 1)\big),$$

and also all the edges between

$$\mathcal{S}_{G_0}(i, m + 1) \cap \mathcal{B}_{G_0}(i', m + 1) \quad \text{and} \quad \mathcal{S}_{G_0}(i, m + 2)$$

in G_0, G, and G', as well as all the edges between

$$\mathcal{S}_{G_0}(i', m + 1) \cap \mathcal{B}_{G_0}(i, m + 1) \quad \text{and} \quad \mathcal{S}_{G_0}(i', m + 2)$$

in G_0, G, and G'. Note that under this conditioning, the quantities $\zeta_s(i, i')$ and $\zeta'_s(i, i')$ defined in (32) are deterministic. Suppose that the instance we condition on satisfies the following statements for fixed constants $K, K', \kappa > 0$:

(C1) $G_0\big(\mathcal{B}_{G_0}(i, m + 1) \cup \mathcal{B}_{G_0}(i', m + 1)\big)$ is a tree or a forest of two trees;
(C2) $|\mathcal{B}_{G_0}(i, m + 1)| + |\mathcal{B}_{G_0}(i', m + 1)| \le n^{0.1}$;
(C3) (29a), (29b), and (31) hold;
(C4) $\big|\big\{j \in \mathcal{S}_{G_0}(i, m) \cup \mathcal{S}_{G_0}(i', m) : \deg_{G \cap G'}(j) \le 2np/3\big\}\big| \le 2(np/3)^m$;

(C5) $|T_s(i)| \wedge |T'_s(i')| \geq (np/2)^m(1-8\kappa)^m$ for all $s \in \{-1,1\}^m$.

We consider the randomness with respect to the remaining possible edges of the graphs; let $\bar{\mathbb{P}}$ and $\bar{\mathbb{E}}$ denote the conditional probability and expectation respectively. Note that for any vertex $j \in \mathcal{S}_{G_0}(i, m+1) \backslash \mathcal{B}_{G_0}(i', m+1)$, the random variable $\deg_G(j) - 1$ is conditionally Binomial$(\bar{n}, p)$, where $\bar{n}$ is defined in (33). Moreover, these binomial variables are independent across different $j \in \mathcal{S}_{G_0}(i, m+1) \backslash \mathcal{B}_{G_0}(i', m+1)$.

In the following, let $f(i)$ and $\mathsf{v}(i)$ be the signature vector and the variance vector respectively given by $\texttt{VertexSignature}(G, i, m)$ (Algorithm 1), and let $f'(i')$ and $\mathsf{v}'(i')$ be given by $\texttt{VertexSignature}(G', i', m)$. Since i and i' are fixed, we omit the dependency of some quantities on (i, i') in the sequel to ease the notation. For example, we omit the argument (i, i') in the quantities L_s and L'_s defined in (30), and in the quantities ζ_s and ζ'_s defined in (32).

Lemma 6.3 (Entrywise difference between signatures). *For every $s \in J$, we have*

$$f_s(i) - f'_s(i') = Z_s + \Delta_s$$

for a random variable Z_s and a deterministic quantity Δ_s satisfying

- $\bar{\mathbb{E}}[Z_s] = 0$;
- $\mathsf{v}_s(i) + \mathsf{v}'_s(i') - 2n^{0.2}p - p(1-p)\bar{n}\big(|L_s| + |L'_s|\big) \leq \bar{\mathbb{E}}[Z_s^2] \leq \mathsf{v}_s(i) + \mathsf{v}'_s(i')$;
- $\bar{\mathbb{P}}\big\{|Z_s| \geq t\big\} \leq 2\exp\Big(\frac{-t^2/2}{\bar{E}[Z_s^2]+t/3}\Big)$;
- $|\Delta_s| \leq |\zeta_s| + |\zeta'_s| + 2n^{0.2}p$.

Moreover, the random variables Z_s are conditionally independent for different $s \in J$.

Proof Similar to the proof of Lemma 5.9, we start with

$$\begin{aligned} f_s(i) = &\sum_{j \in \mathcal{N}_G(T_s(i)) \cap \mathcal{S}_{G_0}(i, m+1)} \big(\deg_G(j) - 1 - \bar{n}p\big) \\ &+ (\bar{n} - n)p\,\big|\mathcal{N}_G(T_s(i)) \cap \mathcal{S}_{G_0}(i, m+1)\big|. \end{aligned}$$

Furthermore, in view of the partition

$$\mathcal{N}_G(T_s(i)) = \big(\mathcal{N}_G(T_s(i)) \backslash \mathcal{B}_{G_0}(i', m+1)\big) \cup \big(\mathcal{N}_G(T_s(i)) \cap \mathcal{B}_{G_0}(i', m+1)\big),$$

we obtain

$$\begin{aligned} f_s(i) = &\sum_{j \in (\mathcal{N}_G(T_s(i)) \backslash \mathcal{B}_{G_0}(i', m+1)) \cap \mathcal{S}_{G_0}(i, m+1)} \big(\deg_G(j) - 1 - \bar{n}p\big) \\ &+ \zeta_s + (\bar{n} - n)p\,\big|\mathcal{N}_G(T_s(i)) \cap \mathcal{S}_{G_0}(i, m+1)\big| \end{aligned}$$

by the definitions of ζ_s in (32) and L_s in (30). An analogous decomposition holds for $f'_s(i')$. Consequently,

$$f_s(i) - f'_s(i') = Z_s + \Delta_s,$$

where

$$Z_s := \sum_{j \in (\mathcal{N}_G(T_s(i)) \setminus \mathcal{B}_{G_0}(i', m+1)) \cap \mathcal{S}_{G_0}(i, m+1)} \big(\deg_G(j) - 1 - \bar{n}p\big) - \sum_{j \in (\mathcal{N}_{G'}(T'_s(i')) \setminus \mathcal{B}_{G_0}(i, m+1)) \cap \mathcal{S}_{G_0}(i', m+1)} \big(\deg_{G'}(j) - 1 - \bar{n}p\big).$$

and

$$\Delta_s := \zeta_s - \zeta'_s + (\bar{n} - n)p\left(\left|\mathcal{N}_G(T_s(i)) \cap \mathcal{S}_{G_0}(i, m+1)\right| - \left|\mathcal{N}_{G'}(T'_s(i')) \cap \mathcal{S}_{G_0}(i', m+1)\right|\right).$$

For the deterministic quantity Δ_s, we have

$$|\Delta_s| \le |\zeta_s| + |\zeta'_s| + n^{0.1}p \cdot 2n^{0.1} = |\zeta_s| + |\zeta'_s| + 2n^{0.2}p.$$

For the random variable Z_s, it is not hard to see that the two sums in its definition are over disjoint sets, so they are independent. Moreover, each sum is the deviation of a binomial random variable from its mean: namely,

$$\sum_{j \in (\mathcal{N}_G(T_s(i)) \setminus \mathcal{B}_{G_0}(i', m+1)) \cap \mathcal{S}_{G_0}(i, m+1)} \big(\deg_G(j) - 1\big)$$

is Binomial$\big(\bar{n} \cdot \big|(\mathcal{N}_G(T_s(i)) \setminus \mathcal{B}_{G_0}(i', m+1)) \cap \mathcal{S}_{G_0}(i, m+1)\big|, p\big)$, and similarly for the other term. Hence, we obtain $\bar{\mathbb{E}}[Z_s] = 0$ and

$$\begin{aligned}
\bar{\mathbb{E}}[Z_s^2] &= p(1-p)\bar{n}\big|(\mathcal{N}_G(T_s(i)) \setminus \mathcal{B}_{G_0}(i', m+1)) \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\quad + p(1-p)\bar{n}\big|(\mathcal{N}_{G'}(T'_s(i)) \setminus \mathcal{B}_{G_0}(i, m+1)) \cap \mathcal{S}_{G_0}(i', m+1)\big| \\
&= p(1-p)\bar{n}\Big(\big|\mathcal{N}_G(T_s(i)) \cap \mathcal{S}_{G_0}(i, m+1)\big| \\
&\quad + \big|\mathcal{N}_{G'}(T'_s(i')) \cap \mathcal{S}_{G_0}(i', m+1)\big| - |L_s| - |L'_s|\Big) \\
&\ge \mathsf{v}_s(i) + \mathsf{v}'_s(i') - 2n^{0.2}p - p(1-p)\bar{n}\big(|L_s| + |L'_s|\big)
\end{aligned}$$

by the definitions of $\mathsf{v}_s(i)$ and $\mathsf{v}'_s(i')$ in Algorithm 1, Condition (C2), and the fact $n - \bar{n} \le n^{0.1}$. It is also obvious that $\bar{\mathbb{E}}[Z_s^2] \le v_s(i) + v'_s(i')$. The tail bound for Z_s follows from Bernstein's inequality. The conditional independence of Z_s for different $s \in J$ follows from the disjointness of sets that we sum over in the definition of Z_s. □

Lemma 6.4 (Difference between signatures). *For any constants $C, D, K, K' > 0$, there exist constants $n_0, R, \kappa > 0$ with the following property. Suppose that*

$$n \ge n_0, \qquad \log n \le nq \le n^{\frac{1}{R \log\log n}}, \qquad \alpha \in (0, \kappa), \qquad 3 \le m \le C \log\log n,$$
$$w \ge \lfloor (\log n)^5 \rfloor.$$

Moreover, suppose that Conditions (C1) to (C5) hold with constants $K, K', \kappa > 0$ and a subset J with $|J| = 2w$. Then it holds with conditional probability at least $1 - n^{-D}$ that

$$\sum_{s\in J} \frac{\big(f_s(i) - f'_s(i')\big)^2}{\mathsf{v}_s(i) + \mathsf{v}'_s(i')} \geq 2w\Big(1 - \frac{1}{(\log n)^{0.9}}\Big).$$

Proof Let

$$I := \Big\{ j \in \mathcal{S}_{G_0}(i, m) : \deg_{G\cap G'}(j) \leq \frac{2}{3}np(1-\alpha)\Big\}.$$

By Conditions (C4) and (C5), we have that for any $s \in \{-1, 1\}^m$,

$$|I| \leq 2(nq/3)^m \leq \frac{1}{2}(np/2)^m(1-8\kappa)^m \leq \frac{1}{2}\big(|T_s(i)| \wedge |T'_s(i')|\big)$$

if $\kappa, \alpha \in (0, 0.01)$ and $m \geq 3$. Since $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ is a tree, it is easy to see that

$$\begin{aligned}\mathsf{v}_s(i) &= np(1-p)\big|\mathcal{N}_G(T_s(i)) \cap \mathcal{S}_G(i, m+1)\big| \geq np(1-p) \sum_{j\in T_s(i)\setminus I} \big(\deg_G(j) - 1\big) \\ &\geq np(1-p) \cdot \frac{1}{2}(np/2)^m(1-8\kappa)^m \cdot \frac{1}{2}np = (1-p)(np/2)^{m+2}(1-8\kappa)^m.\end{aligned}$$

The same estimate also holds for $\mathsf{v}'_s(i')$.

We will apply Lemma 5.10 with

$$X_s := \frac{f_s(i) - f'_s(i')}{\sqrt{\mathsf{v}_s(i) + \mathsf{v}'_s(i')}} = \frac{Z_s + \Delta_s}{\sqrt{\mathsf{v}_s(i) + \mathsf{v}'_s(i')}},$$

where Z_s and Δ_s satisfy the conclusion of Lemma 6.3. Towards that end, let us first establish some estimates for the mean and variance of X_s. Using Lemma 6.3, (31), and the above estimates for $\mathsf{v}_s(i)$ and $\mathsf{v}'_s(i')$, we obtain that for all $s \in J$,

$$\big|\bar{\mathbb{E}}[X_s]\big| = \frac{|\Delta_s|}{\sqrt{\mathsf{v}_s(i) + \mathsf{v}'_s(i')}} \leq \frac{2K'\frac{(np)^{m/2+1}}{2^{m/2}}\sqrt{\log n} + 2n^{0.2}p}{\sqrt{2(1-p)(np/2)^{m+2}(1-8\kappa)^m}} \leq \frac{5K'\sqrt{\log n}}{(1-8\kappa)^{m/2}} \tag{34}$$

and

$$\begin{aligned}\mathrm{Var}(X_s) = \frac{\bar{\mathbb{E}}[Z_s^2]}{\mathsf{v}_s(i) + \mathsf{v}'_s(i')} &\geq \frac{\mathsf{v}_s(i) + \mathsf{v}'_s(i') - 2n^{0.2}p - p(1-p)\bar{n}\big(|L_s| + |L'_s|\big)}{\mathsf{v}_s(i) + \mathsf{v}'_s(i')} \\ &\geq 1 - \frac{2n^{0.2}p + p(1-p)\bar{n}\big(|L_s| + |L'_s|\big)}{2(1-p)(np/2)^{m+2}(1-8\kappa)^m}.\end{aligned}$$

Lemma 6.3 also gives $\mathrm{Var}(X_s) = \frac{\bar{\mathbb{E}}[Z_s^2]}{\mathrm{v}_s(i)+\mathrm{v}'_s(i')} \le 1$ and $\bar{\mathbb{P}}\{|Z_s| \ge t\} \le 2\exp\left(\frac{-t^2/2}{\bar{E}[Z_s^2]+t/3}\right)$. It follows that $\bar{\mathbb{P}}\{|X_s - \bar{\mathbb{E}}[X_s]| \ge t\} \le 2\exp\left(\frac{-t^2/2}{1+t/3}\right)$. Therefore, Lemma 5.10 can be applied to show that, with conditional probability at least $1 - n^{-D}$,

$$\sum_{s\in J} X_s^2 \ge \sum_{s\in J} \bar{\mathbb{E}}[X_s^2] - C_1\sqrt{w}(\log n)^{3/2} - C_1\Big(\max_{s\in J}\big|\bar{\mathbb{E}}[X_s]\big|\Big)\big(\sqrt{w\log n} + \log n\big)$$

for a constant $C_1 > 0$ depending on D, where we recall $|J| = 2w$. Moreover, by the above lower bound on $\mathrm{Var}(X_s) \le \bar{\mathbb{E}}[X_s^2]$ and (29b) assumed in Condition (C3), we have

$$\begin{aligned}
\sum_{s\in J} \bar{\mathbb{E}}[X_s^2] &\ge 2w - \sum_{s\in J} \frac{2n^{0.2}p + p(1-p)\bar{n}\big(|L_s| + |L'_s|\big)}{2(1-p)(np/2)^{m+2}(1-8\kappa)^m} \\
&\ge 2w - n^{-0.7} - \frac{p(1-p)\bar{n}}{2(1-p)(np/2)^{m+2}(1-8\kappa)^m} \\
&\quad \cdot K\frac{(np)^{m+1}}{2^m}\Big(\frac{w}{np(1-\alpha)^m} + \sqrt{w\log n}\Big) \\
&\ge 2w\left(1 - \frac{K_2}{(1-9\kappa)^m \log n}\right)
\end{aligned}$$

for a constant $K_2 > 0$ depending on K, if $w \ge (\log n)^3$, $\log n \le nq \le n^{\frac{1}{K\log\log n}}$, $\alpha < \kappa$, and $n \ge n_0$. The above two bounds together with (34) imply that

$$\begin{aligned}
\sum_{s\in J} X_s^2 &\ge 2w\left(1 - \frac{K_2}{(1-9\kappa)^m \log n}\right) - C_1\sqrt{w}(\log n)^{3/2} \\
&\quad - C_1\frac{5K'\sqrt{\log n}}{(1-8\kappa)^{m/2}}\big(\sqrt{w\log n} + \log n\big) \\
&\ge 2w\left(1 - \frac{K_3}{(1-9\kappa)^m \log n}\right)
\end{aligned}$$

if $w \ge \lfloor(\log n)^5\rfloor$. Finally, if $m \le C\log\log n$ and $\kappa > 0$ is sufficiently small depending on C, then

$$\frac{K_3}{(1-9\kappa)^m \log n} \le \frac{1}{(\log n)^{0.9}},$$

so the proof is complete. □

6.3 Conclusion

We summarize the result of this section in the following proposition.

Proposition 6.5 (Difference between signatures of typical wrong pairs). *For any constants $C, D > 0$, there exist constants $n_0, R, \alpha_0, c > 0$ with the following property. Let J be a uniform random subset of $\{-1,1\}^s$ of cardinality $2w$ for $w \in \mathbb{N}$. Suppose that*

$$n \geq n_0, \qquad \log n \leq np(1-\alpha) \leq n^{\frac{1}{R \log\log n}}, \qquad \alpha \in (0, \alpha_0),$$
$$3 \leq m \leq C \log\log n, \qquad w \geq \lfloor (\log n)^5 \rfloor.$$

Then with probability at least $1 - n^{-D}$, there is a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that for any distinct $i, i' \in \mathcal{I}$,

$$\sum_{s \in J} \frac{\big(f_s(i) - f'_s(i')\big)^2}{v_s(i) + v'_s(i')} \geq 2w\Big(1 - \frac{1}{(\log n)^{0.9}}\Big). \tag{35}$$

Proof We first claim that, with probability at least $1 - n^{-D-1}$, there is a subset $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq n - n^{1-c}$ such that Conditions (C1) to (C5) hold for all distinct $i, i' \in \mathcal{I}$ with constants $K, K', \kappa > 0$. To be more precise, K and K' will depend on C and D, κ is to be chosen, and c depends on C, D, and κ. The proof is very similar to that of Proposition 5.12, so we only provide a sketch and point out the differences.

1. If $\mathrm{dist}_{G_0}(i, i') \leq 2m+2$, then Condition (C1) is weaker than that $G_0\big(\mathcal{B}_{G_0}(i, 3m+3)\big)$ is a tree. If $\mathrm{dist}_{G_0}(i, i') > 2m+2$, then Condition (C1) is saying that $G_0\big(\mathcal{B}_{G_0}(i, m+1)\big)$ and $G_0\big(\mathcal{B}_{G_0}(i', m+1)\big)$ are both trees. In either case, we can show that the neighborhoods of most vertices are trees with high probability as before.
2. Condition (C2) is essentially the same as Condition (B2) up to a constant.
3. Condition (C3) is a direct consequence of Lemma 6.2.
4. Condition (C4) is essentially the same as Condition (B6) up to constants.
5. Condition (C5) is a consequence of Condition (B7) since $|T_s(i)| \geq |T_s(i) \cap T'_s(i)|$.

Therefore, if κ is chosen according to Lemma 6.4, we obtain (35) for any distinct vertices $i, i' \in \mathcal{I}$ with probability at least $1 - n^{-D-2}$. □

Proof of Theorem 2.1 Note that Algorithm 1 is equivariant with respect to the permutation π in the sense that `VertexSignature`$(G^\pi, \pi(i), m)$ and `VertexSignature`(G, i, m) have the same output. Therefore, we may assume without loss of generality that π is the identity. With the choice $m = \lceil 22 \log\log n \rceil$ and $w = \lfloor (\log n)^5 \rfloor$ in Algorithm 2, it is easy to check that the assumptions of Propositions 5.12 and 6.5 are satisfied. Therefore, these two propositions together yield a desired subset $\mathcal{I} \subset [n]$ such that $\sum_{s \in J} \frac{(f_s(i) - f'_s(i))^2}{v_s(i) + v'_s(i)} < 2w\big(1 - \frac{1}{\sqrt{\log n}}\big)$ and $\sum_{s \in J} \frac{(f_s(i) - f'_s(i'))^2}{v_s(i) + v'_s(i')} > 2w\big(1 - \frac{1}{\sqrt{\log n}}\big)$ for distinct $i, i' \in \mathcal{I}$. The theorem follows immediately. □

7 Construction of an exact matching

This section is devoted to developing a procedure that refines an approximate matching to obtain an exact matching. Before proving the main result, we establish some auxiliary statements.

Lemma 7.1 (An elementary decoupling). *Let $M > 0$ be a parameter, let Γ be a fixed graph on $[n]$, and let Q, W be two (possibly intersecting) subsets of vertices of Γ such that*

$$|\mathcal{N}_\Gamma(i) \cap W| \geq M \quad \text{for all } i \in Q.$$

Then there are subsets $Q' \subset Q$ and $W' \subset W$ such that $Q' \cap W' = \varnothing$, $|Q'| \geq |Q|/5$, and

$$|\mathcal{N}_\Gamma(i) \cap W'| \geq M/2 \quad \text{for all } i \in Q'.$$

Proof If $|Q \backslash W| \geq |Q|/5$, then there is nothing to prove, so we can assume that $|Q \cap W| > 4|Q|/5$. Let $\hat{Q}$ be a uniform random subset of $Q \cap W$. Consider random disjoint sets $\hat{Q}$ and $\hat{W} := (W \backslash Q) \cup ((Q \cap W) \backslash \hat{Q})$. Fix $i \in Q \cap W$. Note that each $j \in \mathcal{N}_\Gamma(i) \cap W$ belongs to $\mathcal{N}_\Gamma(i) \cap \hat{W}$ with probability at least $1/2$, and $|\mathcal{N}_\Gamma(i) \cap W| \geq M$ by assumption, so the variable $|\mathcal{N}_\Gamma(i) \cap \hat{W}|$ dominates Binomial$(M, 1/2)$ stochastically. Since this is still true conditional on $i \in \hat{Q}$, we have

$$\begin{aligned}\mathbb{P}\big\{i \in \hat{Q},\ |\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2\big\} &= \mathbb{P}\big\{i \in \hat{Q}\big\} \cdot \mathbb{P}\big\{|\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2 \mid i \in \hat{Q}\big\} \\ &\geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}\big|\big\{i \in \hat{Q} : |\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2\big\}\big| &= \sum_{i \in Q \cap W} \mathbb{P}\big\{i \in \hat{Q},\ |\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2\big\} \\ &\geq \frac{|Q \cap W|}{4} \geq \frac{|Q|}{5}.\end{aligned}$$

Therefore, there is a realization of $(\hat{Q}, \hat{W})$ such that $\big|\big\{i \in \hat{Q} : |\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2\big\}\big| \geq |Q|/5$. It then suffices to take $W' = \hat{W}$ and $Q' = \big\{i \in \hat{Q} : |\mathcal{N}_\Gamma(i) \cap \hat{W}| \geq M/2\big\}$. □

For any subset $I \subset [n]$, let $I^c := [n] \backslash I$.

Lemma 7.2 (Growing a subset of vertices). *For any $\delta' \in (0, 1/2]$, there are $n_0, c > 0$ depending on δ' with the following property. Assume $n > n_0$, and that $rn \geq \log n$. Let Γ be a $G(n, r)$ graph, and let I be a random subset of $[n]$ (possibly depending on Γ), with $|I| \geq n - \delta' n/6$. Define a random subset of vertices*

$$\tilde{I} := \Big\{i \in [n] :\ |\mathcal{N}_\Gamma(i) \cap I^c| < \delta' rn\Big\}.$$

Then we have

$$\mathbb{P}\Big\{|\tilde{I}^c| \le \frac{1}{4}|I^c|\Big\} \ge 1 - \exp(-c\,rn\log n).$$

Proof Given the assumptions on I, by considering sets $W = I^c$ and $Q \subset \tilde{I}^c$, we obtain

$$\begin{aligned}&\big\{|\tilde{I}^c| > |I^c|/4\big\} \subset\\ &\mathcal{E} := \Big\{\exists\ Q, W \subset [n]:\ |W| \le \delta' n/6,\ |Q|\\ &\quad = \lceil |W|/4\rceil \ne 0,\ |\mathcal{N}_\Gamma(i) \cap W| \ge \delta' rn \text{ for all } i \in Q\Big\}.\end{aligned}$$

Further, according to Lemma 7.1, $\mathcal{E}$ is contained in the event

$$\begin{aligned}\mathcal{E}' := \Big\{&\exists\ Q', W' \subset [n]:\ \ |W'| \le \delta' n/6,\ |Q'| \ge \lceil\lceil |W'|/4\rceil/5\rceil \ne 0,\ Q' \cap W' = \varnothing,\\ &|\mathcal{N}_\Gamma(i) \cap W'| \ge \delta' rn/2 \text{ for all } i \in Q'\Big\}.\end{aligned}$$

We estimate the probability of $\mathcal{E}'$ by taking the union bound over all possible realizations Q' and W' (note that necessarily $|W'| \ge \delta' rn/2$ in the event description). Observe that for any disjoint fixed subsets Q' and W', the binomial variables $|\mathcal{N}_\Gamma(i) \cap W'|$, $i \in Q'$, are mutually independent (this is the reason for applying the decoupling lemma, Lemma 7.1). We then get an upper bound

$$\mathbb{P}\big\{|\tilde{I}^c| > |I^c|/4\big\} \le \sum_{w=\lceil \delta' rn/2\rceil}^{\lfloor \delta' n/6\rfloor} \binom{n}{w} \sum_{k=\lceil\lceil w/4\rceil/5\rceil}^{n} \binom{n}{k} \mathbb{P}\Big\{\sum_{i=1}^{w} b_i \ge \delta' rn/2\Big\}^k,$$

where $b_1, \dots, b_w$ are i.i.d. Bernoulli(r) variables.

Applying Chernoff's inequality (first estimate in Lemma 4.1), we get

$$\mathbb{P}\Big\{\sum_{i=1}^{w} b_i \ge \delta' rn/2\Big\} \le \Big(\frac{rw}{\delta' rn}\Big)^{c_1 \delta' rn},$$

for some universal constant $c_1 > 0$ and all $w \le \delta' n/6$ (assuming that n is large enough). Thus,

$$\mathbb{P}\big\{|\tilde{I}^c| > |I^c|/4\big\} \le \sum_{w=\lceil \delta' rn/2\rceil}^{\lfloor \delta' n/6\rfloor} \sum_{k=\lceil\lceil w/4\rceil/5\rceil}^{n} \Big(\frac{en}{w}\Big)^w \Big(\frac{en}{k}\Big)^k \Big(\frac{w}{\delta' n}\Big)^{c_1 \delta' rnk}.$$

A straightforward computation completes the proof. □

Lemma 7.3 (Number of neighbors). *For any $\varepsilon \in (0,1]$ and $\alpha \in (0,\varepsilon/4]$, there is $n_0 > 0$ depending only on ε with the following property. Let $n > n_0$ and $q \in (0,1)$. Assume that $p := (1-\alpha)q$ satisfies $pn \geq (1+\varepsilon)\log n$. Let G_0, G, and G' be as before. Then with probability at least $1 - \exp(-\varepsilon pn/8)$,*

$$|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)| \geq \varepsilon^2 pn/256 \quad \text{for all } i \in [n].$$

Proof Pick any vertex $i \in [n]$. The variable $|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)|$ is Binomial$(n-1,(1-\alpha)p)$, so, applying Chernoff's inequality (second estimate in Lemma 4.1), we get

$$\begin{aligned}\mathbb{P}\big\{|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)| < u(1-\alpha)p(n-1)\big\} \\ \leq \exp(-(1-\alpha)p(n-1)) \cdot (e/u)^{u(1-\alpha)p(n-1)}\end{aligned}$$

for every $u \in (0,1)$. Thus, assuming that n is large enough, we have

$$\begin{aligned}\mathbb{P}\big\{|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)| < \varepsilon^2 pn/256\big\} &\leq \exp(-(1-\alpha)p(n-1))\exp(\varepsilon pn/16) \\ &\leq \exp(-(1-3\varepsilon/8)pn).\end{aligned}$$

Taking the union bound over $i \in [n]$, we get

$$\mathbb{P}\big\{|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)| \geq \varepsilon^2 pn/256 \text{ for all } i \in [n]\big\} \geq 1 - n\exp(-(1-3\varepsilon/8)pn).$$

It remains to use the assumption $pn \geq (1+\varepsilon)\log n$ to get the result. □

The next lemma is just a restatement of Lemma 4.6 in a more specific context:

Lemma 7.4 (Number of common neighbors of distinct vertices). *For any $\delta'' > 0$ there are $n_0'' \in \mathbb{N}$ and $c'' > 0$ depending on δ'' with the following property. Assume that $n > n_0''$ and that $p \in (0,1/2]$ and $\alpha \in (0,1/2]$ satisfy $pn \geq \log n$ and $4pn\log n \leq \sqrt{n}$. Let G_0, G, and G' be as before. Then with probability at least $1 - \exp(-c''pn\log n)$, we have*

$$|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i')| \leq \delta'' pn \quad \text{for all } i, i' \in [n],\ i \neq i'.$$

Proof The result follows immediately by applying Lemma 4.6 with $\Gamma = G_0$ and $J = \{i, i'\}$, together with a union bound over distinct $i, i' \in [n]$. □

Proposition 7.5 (Improving a partial matching). *For every $\varepsilon \in (0,1]$, there exists $n_0 > 0$ and $\kappa \in (0,1)$ depending on ε with the following property. Assume that*

$$n \geq n_0, \qquad (1+\varepsilon)\log n \leq pn \leq \frac{\sqrt{n}}{4\log n}, \qquad \alpha \in (0,\varepsilon/4].$$

Let the graphs G and G' be as before. Assume that $\mathcal{J}$ is a random subset of $[n]$ (possibly depending on G and G'), and that $g : \mathcal{J} \to [n]$ is a random injective mapping (again,

possibly depending on G and G'). Let

$$\mathcal{E} := \big\{\big|\{i \in \mathcal{J} :\ g(i) = i\}\big| \geq n - \kappa n\big\}.$$

Define a random subset $\tilde{\mathcal{J}} \subset [n]$ and a random injective mapping $\tilde{g} : \tilde{\mathcal{J}} \to [n]$ as follows: For every $i \in [n]$, i is in the set $\tilde{\mathcal{J}}$ if and only if there is a (unique) vertex $i' \in [n]$ such that

- $|g(\mathcal{N}_G(i) \cap \mathcal{J}) \cap \mathcal{N}_{G'}(i')| \geq \varepsilon^2 pn/512$;
- $|g(\mathcal{N}_G(i) \cap \mathcal{J}) \cap \mathcal{N}_{G'}(j)| < \varepsilon^2 pn/512$ *for all* $j \in [n]\backslash\{i'\}$;
- $|g(\mathcal{N}_G(j) \cap \mathcal{J}) \cap \mathcal{N}_{G'}(i')| < \varepsilon^2 pn/512$ *for all* $j \in [n]\backslash\{i\}$.

We then set $\tilde{g}(i) := i'$ for any such pair of vertices i and i'.

Then with probability at least $\mathbb{P}(\mathcal{E}) - \exp(-\varepsilon pn/9)$,

$$\big|\big\{i \in \tilde{\mathcal{J}} :\ \tilde{g}(i) = i\big\}\big| \geq \frac{1}{2}n + \frac{1}{2}\big|\big\{i \in \mathcal{J} :\ g(i) = i\big\}\big|.$$

Proof Define random sets

$$\begin{aligned} I &:= \{j \in \mathcal{J} :\ g(j) = j\}, \\ \tilde{I} &:= \big\{i \in [n] :\ \big|\mathcal{N}_G(i) \cap I^c\big| < 2^{-10}\varepsilon^2 pn\big\}, \\ \tilde{I}' &:= \big\{i \in [n] :\ \big|\mathcal{N}_{G'}(i) \cap I^c\big| < 2^{-10}\varepsilon^2 pn\big\}, \end{aligned}$$

and consider the event

$$\begin{aligned} \mathcal{E}' := \Big\{&\big|\tilde{I}^c\big| \vee \big|(\tilde{I}')^c\big| \leq \frac{1}{4}|I^c|, \quad \big|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)\big| \geq \varepsilon^2 pn/256 \quad \text{for all } i \in [n], \ \text{ and} \\ &\big|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i')\big| < 2^{-10}\varepsilon^2 pn \quad \text{for all } i, i' \in [n],\ i \neq i'\Big\}. \end{aligned}$$

If we choose $\kappa > 0$ sufficiently small and n_0 sufficiently large depending on ε, then, assuming $n > n_0$, the event $\mathcal{E}' \cap \mathcal{E}$ has probability at least

$$\mathbb{P}(\mathcal{E}) - \exp(-\varepsilon pn/9),$$

by combining Lemmas 7.2, 7.3, and 7.4. We claim that everywhere on $\mathcal{E}' \cap \mathcal{E}$, the set $\tilde{\mathcal{J}}$ and the mapping $\tilde{g}$ satisfy the conclusions of the proposition.

Condition on any realization of $G, G', \mathcal{J}, g$ such that the event $\mathcal{E}' \cap \mathcal{E}$ holds. Pick any vertex $i \in \tilde{I} \cap \tilde{I}'$. By the definition of $\mathcal{E}' \cap \mathcal{E}$, we have

$$\begin{aligned} \big|g(\mathcal{N}_G(i) \cap \mathcal{J}) \cap \mathcal{N}_{G'}(i)\big| &\geq \big|\mathcal{N}_G(i) \cap I \cap \mathcal{N}_{G'}(i)\big| \\ &\geq \big|\mathcal{N}_G(i) \cap \mathcal{N}_{G'}(i)\big| - \big|\mathcal{N}_G(i) \cap I^c\big| \\ &\geq \varepsilon^2 pn/256 - 2^{-10}\varepsilon^2 pn \geq \varepsilon^2 pn/512. \end{aligned}$$

On the other hand, for every $i' \in [n]\backslash\{i\}$ we have

$$\begin{aligned}\left|g(\mathcal{N}_G(i)\cap\mathcal{J})\cap\mathcal{N}_{G'}(i')\right| &\le \left|\mathcal{N}_G(i)\cap I\cap\mathcal{N}_{G'}(i')\right| + \left|g(\mathcal{N}_G(i)\cap(\mathcal{J}\backslash I))\cap\mathcal{N}_{G'}(i')\right| \\ &\le \left|\mathcal{N}_G(i)\cap\mathcal{N}_{G'}(i')\right| + \left|\mathcal{N}_G(i)\cap I^c\right| \\ &< 2^{-10}\varepsilon^2 pn + 2^{-10}\varepsilon^2 pn = \varepsilon^2 pn/512,\end{aligned}$$

and, similarly,

$$\begin{aligned}\left|g(\mathcal{N}_G(i')\cap\mathcal{J})\cap\mathcal{N}_{G'}(i)\right| &\le \left|\mathcal{N}_G(i')\cap I\cap\mathcal{N}_{G'}(i)\right| + \left|g(\mathcal{N}_G(i')\cap(\mathcal{J}\backslash I))\cap\mathcal{N}_{G'}(i)\right| \\ &\le \left|\mathcal{N}_G(i')\cap\mathcal{N}_{G'}(i)\right| + \left|I^c\cap\mathcal{N}_{G'}(i)\right| \\ &< 2^{-10}\varepsilon^2 pn + 2^{-10}\varepsilon^2 pn = \varepsilon^2 pn/512.\end{aligned}$$

Thus, $\tilde{I}\cap\tilde{I}'\subset\tilde{\mathcal{J}}$ and $\tilde{g}(i)=i$ for all $i\in\tilde{I}\cap\tilde{I}'$. Moreover, by the first condition in $\mathcal{E}'$, we have

$$\left|\tilde{I}\cap\tilde{I}'\right| \ge n - \left|\tilde{I}^c\right| - \left|(\tilde{I}')^c\right| \ge n - \frac{1}{2}|I^c| = \frac{1}{2}n + \frac{1}{2}|I|,$$

so the result follows. □

Proof of Theorem 2.4 In short, the theorem follows by applying Proposition 7.5 iteratively.

To be more precise, first note that whenever we set $\pi_\ell(i')=i$ in Algorithm 4, it is impossible to have $\pi_\ell(j)=i$ for $j\neq i'$ or $\pi_\ell(i')=j$ for $j\neq i$ thanks to the three conditions. As a result, π_ℓ is a well-defined injective function between subsets of $[n]$ after the loop through $i=1,\dots,n$, and so π_ℓ can be extended to a permutation on $[n]$.

Next, for any $\ell\in\left[\lceil\log_2 n\rceil\right]$ and $i\in[n]$, we have $\pi_{\ell-1}^{-1}\big(\mathcal{N}_{G^\pi}(i)\big) = \pi_{\ell-1}^{-1}\circ\pi\big(\mathcal{N}_G(\pi^{-1}(i))\big)$. Denote $g_\ell=\pi_\ell^{-1}\circ\pi$. Therefore, when setting $\pi_\ell(i')=i$ in Algorithm 4, we are defining $g_\ell\big(\pi^{-1}(i)\big)=i'$ if the conditions

- $\left|g_{\ell-1}\big(\mathcal{N}_G(\pi^{-1}(i))\big)\cap\mathcal{N}_{G'}(i')\right| \ge \varepsilon^2 pn/512$,
- $\left|g_{\ell-1}\big(\mathcal{N}_G(\pi^{-1}(i))\big)\cap\mathcal{N}_{G'}(j)\right| < \varepsilon^2 pn/512$ for all $j\in[n]\backslash\{i'\}$, and
- $\left|g_{\ell-1}\big(\mathcal{N}_G(j)\big)\cap\mathcal{N}_{G'}(i')\right| < \varepsilon^2 pn/512$ for all $j\in[n]\backslash\{\pi^{-1}(i)\}$

are satisfied. Replacing $\pi^{-1}(i)$ by i in the last statement (which is simply a change of notation as i varies over $[n]$), we see that Proposition 7.5 can be applied with $\mathcal{J}=[n]$ to yield the following: With probability at least

$$\mathbb{P}\left\{\left|\{i\in[n]: g_{\ell-1}(i)\neq i\}\right| \le \frac{\kappa n}{2^{\ell-1}}\right\} - \exp(-\varepsilon pn/9),$$

we have

$$\left|\left\{i\in[n]: g_\ell(i)\neq i\right\}\right| \le \frac{\kappa n}{2^\ell}.$$

To conclude, note that

$$\big|\big\{i \in [n] : g_\ell(i) \neq i\big\}\big| = \big|\big\{i \in [n] : \pi_\ell^{-1} \circ \pi(i) \neq i\big\}\big| = \big|\big\{i \in [n] : \pi_\ell(i) \neq \pi(i)\big\}\big|.$$

Since $\pi_0 = \hat{\pi}$ and $\pi_{\lceil \log_2 n \rceil} = \tilde{\pi}$, applying the above argument iteratively for $\ell = 1, \ldots, \lceil \log_2 n \rceil$ gives that, with probability at least

$$\mathbb{P}\big\{|\{i \in [n] : \hat{\pi}(i) \neq \pi(i)\}| \leq \kappa n\big\} - \lceil \log_2 n \rceil \cdot \exp(-\varepsilon pn/9),$$

we have

$$\big|\big\{i \in [n] : \tilde{\pi}(i) \neq \pi(i)\big\}\big| \leq \frac{\kappa n}{2^{\lceil \log_2 n \rceil}} < 1,$$

that is, $\tilde{\pi} = \pi$. Since $\lceil \log_2 n \rceil \cdot \exp(-\varepsilon pn/9) \leq \exp(-\varepsilon pn/10)$ by the assumptions on the parameters, this completes the proof. □

References

1. Babai, L.: Graph isomorphism in quasipolynomial time. In: Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing, pp. 684–697 (2016)
2. Barak, B., Chou, C.-N., Lei, Z., Schramm, T., Sheng, Y.: (Nearly) efficient algorithms for the graph matching problem on correlated random graphs. In: Advances in Neural Information Processing Systems, pp. 9190–9198 (2019)
3. Burkard, R.E., Cela, E., Pardalos, P.M., Pitsoulis, L.S.: The quadratic assignment problem. In: Handbook of Combinatorial Optimization, pp. 1713–1809. Springer (1998)
4. Babai, L., Erdös, P., Selkow, S.M.: Random graph isomorphism. SIAM J. Comput. **9**(3), 628–635 (1980)
5. Bollobás, B.: Distinguishing vertices of random graphs. North-Holland Math. Stud. **62**, 33–49 (1982)
6. Bozorg, M., Salehkaleybar, S., Hashemi, M.: Seedless graph matching via tail of degree distribution for correlated Erdos–Renyi graphs (2019). arXiv preprint arXiv:1907.06334
7. Cullina, D., Kiyavash, N.: Improved achievability and converse bounds for Erdös–Rényi graph matching. In: Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, pp. 63–72. ACM (2016)
8. Cullina, D., Kiyavash, N., Mittal, P., Poor, H.V.: Partial recovery of erdős-rényi graph alignment via k-core alignment. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems, vol. 3, no. 3, pp. 1–21 (2019)
9. Czajka, T., Pandurangan, G.: Improved random graph isomorphism. J. Discrete Algorithms **6**(1), 85–92 (2008)
10. Dai, O.E., Cullina, D., Kiyavash, N., Grossglauser, M.: Analysis of a canonical labeling algorithm for the alignment of correlated Erdos–Rényi graphs. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems, vol. 3, no. 2, pp. 1–25 (2019)
11. Ding, J., Ma, Z., Yihong, W., Jiaming, X.: Efficient random graph matching via degree profiles. Probab. Theory Relat. Fields **179**(1), 29–115 (2021)
12. Fan, Z., Mao, C., Wu, Y., Xu, J.: Spectral graph matching and regularized quadratic relaxations II: Erdős–Rényi graphs and universality (2019). arXiv preprint arXiv:1907.08883
13. Fan, Z., Mao, C., Wu, Y., Xu, J.: Spectral graph matching and regularized quadratic relaxations: algorithm and theory. In: International Conference on Machine Learning, pp. 2985–2995. PMLR (2020)

14. Feizi, S., Quon, G., Recamonde-Mendoza, M., Médard, M., Kellis, M., Jadbabaie, A.: Spectral alignment of networks (2016). arXiv preprint arXiv:1602.04181
15. Ganassali, L., Massoulié, L.: From tree matching to sparse graph alignment. In: Proceedings of Thirty Third Conference on Learning Theory, Volume 125 of Proceedings of Machine Learning Research, pp. 1633–1665. PMLR (2020)
16. Ganassali, L., Massoulié, L., Lelarge, M.: Correlation detection in trees for partial graph alignment (2021). arXiv preprint arXiv:2107.07623
17. Kazemi, E., Hassani, S.H., Grossglauser, M.: Growing a graph matching from a handful of seeds. Proc. VLDB Endowment **8**(10), 1010–1021 (2015)
18. Lyzinski, V., Fishkind, D.E., Priebe, C.E.: Seeded graph matching for correlated Erdös–Rényi graphs. J. Mach. Learn. Res. **15**(1), 3513–3540 (2014)
19. Lubars, J., Srikant, R.: Correcting the output of approximate graph matching algorithms. In: IEEE INFOCOM 2018-IEEE Conference on Computer Communications, pp. 1745–1753. IEEE (2018)
20. Mao, C., Rudelson, M., Tikhomirov, K.: Random graph matching with improved noise robustness. In: Proceedings of Thirty Fourth Conference on Learning Theory, Volume 134 of Proceedings of Machine Learning Research, pp. 3296–3329. PMLR (2021)
21. Mossel, E., Jiaming, X.: Seeded graph matching via large neighborhood statistics. Random Struct. Algorithms **57**(3), 570–611 (2020)
22. Pedarsani, P., 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)
23. Pardalos, P.M., Rendl, F., Wolkowicz, H.: The quadratic assignment problem: a survey and recent developments. In: Proceedings of the DIMACS Workshop on Quadratic Assignment Problems, Volume 16 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 1–42. American Mathematical Society (1994)
24. Shirani, F., Garg, S., Erkip, E.: Seeded graph matching: efficient algorithms and theoretical guarantees. In: 2017 51st Asilomar Conference on Signals, Systems, and Computers, pp. 253–257. IEEE (2017)
25. Vershynin, R.: High-Dimensional Probability, Volume 47 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2018) **(An introduction with applications in data science, With a foreword by Sara van de Geer)**
26. Wu, Y., Xu, J., Yu, S.H.: Settling the sharp reconstruction thresholds of random graph matching (2021). arXiv preprint arXiv:2102.00082
27. Yartseva, L., Grossglauser, M.: On the performance of percolation graph matching. In: Proceedings of the First ACM Conference on Online Social Networks, pp. 119–130. ACM (2013)
28. Yu, L., Xu, J., Lin, X.: Graph matching with partially-correct seeds (2020). arXiv preprint arXiv:2004.03816