Proceedings of Machine Learning Research vol 196:1–25, 2024 37th Annual Conference on Learning Theory

Information-Theoretic Thresholds for the Alignments of Partially Correlated Graphs

Dong Huang HD23@MAILS.TSINGUA.EDU.CN
Center for Statistical Science, Tsinghua University

Xianwen Song SXW20@MAILS.TSINGHUA.EDU.CN
Qiuzhen College, Tsinghua Universiy

Pengkun Yang YANGPENGKUN@TSINGHUA.EDU.CN
Center for Statistical Science, Tsinghua University

Editors: Shipra Agrawal and Aaron Roth

Abstract

This paper studies the problem of recovering the hidden vertex correspondence between two correlated random graphs. We propose the partially correlated Erdős-Rényi graphs model, wherein a pair of induced subgraphs with a certain number are correlated. We investigate the information-theoretic thresholds for recovering the latent correlated subgraphs and the hidden vertex correspondence. We prove that there exists an optimal rate for partial recovery for the number of correlated nodes, above which one can correctly match a fraction of vertices and below which correctly matching any positive fraction is impossible, and we also derive an optimal rate for exact recovery. In the proof of possibility results, we propose correlated functional digraphs, which partition the edges of the intersection graph into two types of components, and bound the error probability by lower-order cumulant generating functions. The proof of impossibility results build upon the generalized Fano's inequality and the recovery thresholds settled in correlated Erdős-Rényi graphs model.

Keywords: Graph alignments, information-theoretic thresholds, Erdős-Rényi random graphs, partial recovery, exact recovery

1. Introduction

Recently, there has been a surge in interest in the problems of detecting graph correlations and the alignments of two correlated graphs. These questions have emerged across various domains. For instance, in social networks, determining the similarity between friendship networks across different platforms has garnered attention (Narayanan and Shmatikov, 2008, 2009).

In the realm of computer vision, where 3-D shapes are often represented as graphs with adjacency matrices, the identification of whether two graphs represent the same object holds significant importance in pattern recognition and image processing (Berg et al., 2005; Cour et al., 2006). In computational biology, the representation of biological networks as graphs aids in understanding and quantifying their correlation (Singh et al., 2008; Vogelstein et al., 2011). Furthermore, in natural language processing, the ontology alignment problem involves representing each sentence as a graph, with nodes denoting words. The task of determining whether a given sentence can be inferred from the text directly relates to graph matching problems (Haghighi et al., 2005). Numerous graph models exist, with the Erdős-Rényi random graph model being a prominent example, as proposed by Paul and Alfréd (1959) and Gilbert (1959):

Definition 1 (Erdős-Rényi graph) *The Erdős-Rényi random graph is the graph on n vertices where each edge connects with probability $0 < p < 1$ independently. Let $\mathcal{G}(n, p)$ denote the distribution of Erdős-Rényi random graphs with n vertices and edge connecting probability p.*

While there are inherent disparities between the Erdős-Rényi random graph model and networks derived from real-world scenarios, comprehensively understanding the Erdős-Rényi graphs remains profoundly significant. This understanding serves as a pivotal step in transitioning from solving detection and matching problems on Erdős-Rényi graphs to addressing challenges inherent in practical applications. The graph alignment problem entails identifying latent vertex correspondences between two graphs based on their structures. Following Pedarsani and Grossglauser (2011), for two random graphs G_1, G_2 with vertex sets $V(G_1), V(G_2)$ and edge sets $E(G_1), E(G_2)$, a typical correlated graph model is correlated Erdős-Rényi random graph model:

Definition 2 (Correlated Erdős-Rényi graphs) *Let π denote a latent bijective mapping from $V(G_1)$ to $V(G_2)$. We say a pair of graphs (G_1, G_2) are correlated Erdős-Rényi graphs if both marginal distributions are $\mathcal{G}(n, p)$ and each pair of edges $(uv, \pi(u)\pi(v))$ for $u, v \in V(G_1)$ follows the correlated bivariate Bernoulli distribution with correlation coefficient ρ.*

Given observations on G_1 and G_2 under the correlated Erdős-Rényi graphs model, the goal is to recover the latent vertex mapping π. To quantify the performance of an estimator $\hat{\pi}$, we consider the following two recovery criterion:

- *Partial recovery*: given a constant $\delta \in (0, 1)$, we say $\hat{\pi}$ succeeds for partial recovery if

$$|\{v \in V(G_1) : \pi(v) = \hat{\pi}(v)\}| \geq \delta |V(G_1)|. \tag{1}$$

- *Exact recovery*: we say $\hat{\pi}$ succeeds for exact recovery if

$$\pi(v) = \hat{\pi}(v), \quad \forall\, v \in V(G_1). \tag{2}$$

The information-theoretic thresholds for partial and exact recoveries of π between two correlated Erdős-Rényi graphs have been extensively studied in the recent literature.

- *Partial Recovery*. Ganassali et al. (2021) presented an impossibility result for partial recovery in the sparse regime characterized by constant average degree and correlation. Hall and Massoulié (2023) showed that $np(p \vee \rho) \gtrsim \log\left(1 + \frac{\rho}{p}\right) \vee 1$ suffices for partial recovery, while $n \gtrsim d(p + \rho - p\rho \| p) \log n$ is necessary, where $d(p\|q)$ denotes the Kullback–Leibler (KL) divergence between Bernoulli distributions with mean p and q, respectively. The recent work Wu et al. (2022) settled the sharp threshold for dense graphs with $\frac{p}{p \vee \rho} = n^{-o(1)}$ and the thresholds within a constant factor for sparse ones with $\frac{p}{p \vee \rho} = n^{-\Omega(1)}$. For the sparse case, Ding and Du (2023b) proved a sharp threshold when $\frac{p}{p \vee \rho} = n^{-\alpha + o(1)}$ for $\alpha \in (0, 1]$.

- *Exact Recovery*. Based on the properties of the intersection graph under a permutation π, (Cullina and Kiyavash, 2016, 2017) showed that the Maximal Likelihood Estimator (MLE) achieves exact recovery and established an information-theoretical lower bound with a gap of $\omega(1)$. The results are sharpened by Wu et al. (2022) where the sharp threshold for exact recovery are derived.

While numerous studies have extensively investigated recovery procedures within the correlated Erdős-Rényi graphs model, it is however imperative to recognize that the signal present in many graph structures from realistic models is often inferior to that within the correlated Erdős-Rényi graph. This discrepancy emerges as many nodes in realistic graphs do not have corresponding nodes in the second correlated graphs. To offer a resolution to this concern, we propose the following model where on part of the nodes from two graphs are correlated.

Definition 3 (Partially correlated Erdős-Rényi graphs) *Let $S^* \subseteq V(G_1)$ be a latent subset of vertices and π^* be a latent injective mapping from S^* to $V(G_2)$. We say a pair of graphs (G_1, G_2) are* partially *correlated Erdős-Rényi graphs if both marginal distributions are $\mathcal{G}(n,p)$ and each pair of edges $(uv, \pi^*(u)\pi^*(v))$ for $u, v \in S^*$ follows the correlated bivariate Bernoulli distribution with correlation coefficient ρ.*

The case $S^* = V(G_1)$ reduces to a pair of correlated Erdős-Rényi graphs in Definition 2. Under the model in Definition 3, given $S^* \subseteq V(G_1)$ and the range of π^* denoted by $T^* \subseteq V(G_2)$, the induced subgraphs $G_1[S^*]$ and $G_2[T^*]$ are correlated Erdős-Rényi graphs on m vertices. Therefore, the model can be equivalently constructed by planting correlated Erdős-Rényi graphs over a pair of independent Erdős-Rényi graphs.

In this paper, we investigate the information-theoretic thresholds for recovering the correlated nodes S^* and the mapping π^*. For notational simplicity, we also refer to the problem as recovering π^* while keeping S^* implicit as the domain of π^*. The success criterion is similar to (1) and (2), where $V(G_1)$ shall be replaced by S^*. However, due to the potential inconsistency between the domain of π^* and the estimator $\hat{\pi} : \hat{S} \mapsto V(G_2)$, we define their overlap by:

$$\text{overlap}(\pi^*, \hat{\pi}) \triangleq \frac{|v \in S^* \cap \hat{S} : \pi^*(v) = \hat{\pi}(v)|}{|S^*|}. \tag{3}$$

With the notion of overlap, the success criterion is given by

- *Partial recovery*: $\hat{\pi}$ succeeds if $\text{overlap}(\pi^*, \hat{\pi}) \geq \delta$ for a given constant $\delta \in (0, 1)$;
- *Exact recovery*: $\hat{\pi}$ succeeds if $\text{overlap}(\pi^*, \hat{\pi}) = 1$.

1.1. Main Results

In this subsection, we present the main results of the paper. We first introduce some notations for the presentation of main theorems. Throughout the paper, we assume $0 < \rho \leq 1$, $0 < p \leq \frac{1}{2}$, and the cardinality $|S^*| = m$ is known. We further assume $p \geq \frac{1}{n}$ since otherwise partial recovery is impossible by Wu et al. (2022). For a pair of Bernoulli random variables with means p_1, p_2 and correlation ρ, their bivariate distribution is denoted as $\text{Bern}(p_1, p_2, \rho)$. In our model, a pair of correlated edges $(e, \pi^*(e)) \sim \text{Bern}(p, p, \rho)$. Define $p_{ij} \triangleq \mathbb{P}[e = i, \pi^*(e) = j]$ for $i, j \in \{0, 1\}$. Then

$$p_{11} = p^2 + \rho p(1-p), \quad p_{10} = p_{01} = (1-\rho)p(1-p), \quad p_{00} = (1-p)^2 + \rho p(1-p).$$

For a pair $(e, \pi^*(e))$ both edges present with probability p_{11}, while for $\pi(e) \neq \pi^*(e)$ both e and $\pi(e)$ present with probability p^2. The relative signal strength present in correlated edges is denoted by $\gamma \triangleq \frac{p_{11}}{p^2} - 1 = \frac{\rho(1-p)}{p}$. It turns out that such reparametrization of the correlation coefficient is crucial in determining the fundamental limits of the graph alignment problem.

Let $\mathcal{S}_{n,m}$ denote the set of injective mappings $\pi : S \subseteq V(G_1) \mapsto V(G_2)$ with $|S| = m$. Our goal is to identify the minimum number of correlated nodes m such that recovery of π^* is possible. For the possibility results, we consider the estimator defined as

$$\hat{\pi} = \underset{\pi \in \mathcal{S}_{n,m}}{\operatorname{argmax}} \sum_{u \neq v} \mathbb{1}_{uv \in E(G_1)} \mathbb{1}_{\pi(u)\pi(v) \in E(G_2)}. \tag{4}$$

Next, we introduce our main theorems. Define $\phi(\gamma) \triangleq (1+\gamma)\log(1+\gamma) - \gamma$.

Theorem 4 (Partial recovery) *For any constant $\delta \in (0,1)$, there exists constant $c_1(\delta)$ such that, when $m \geq \frac{c_1(\delta)\log n}{p^2\phi(\gamma)}$, for any $\pi^* \in \mathcal{S}_{n,m}$, the estimator in (4) satisfies*

$$\mathbb{P}\left[\operatorname{overlap}(\pi^*, \hat{\pi}) \geq \delta\right] = 1 - o(1).$$

Furthermore, for any $c \in (0,1)$, there exists $c_2(c,\delta)$ such that, when $m \leq \frac{c_2(c,\delta)\log n}{p^2\phi(\gamma)}$, for any estimator $\hat{\pi}$,

$$\mathbb{P}\left[\operatorname{overlap}(\pi^*, \hat{\pi}) < \delta\right] \geq 1 - c,$$

where π^* is uniformly distributed over $\mathcal{S}_{n,m}$.

The possibility result is presented in the minimax sense, while the impossibility result is under a Bayesian model. Hence, the threshold holds for both minimax and Bayesian risks. Theorem 4 implies, for the purpose of partial recovery, the threshold for the number of correlated nodes m is of the order $\frac{\log n}{p^2\phi(\gamma)}$, beyond which partial recovery is possible and below which partial recovery is impossible. The dependency on the ambient graph order is only logarithmic, while the scale in terms of p and ρ is characterized by $\frac{1}{p^2\phi(\gamma)}$.

Theorem 5 (Exact recovery) *When $m \geq C\left(\frac{\log n}{p^2\phi(\gamma)} \vee \frac{\log(1/(p^2\gamma))}{p^2\gamma}\right)$, where C is a universal constant, for any $\pi^* \in \mathcal{S}_{n,m}$, the estimator in (4) satisfies*

$$\mathbb{P}\left[\operatorname{overlap}(\pi^*, \hat{\pi}) = 1\right] = 1 - o(1).$$

Furthermore, for any $c \in (0,1)$, there exists a constant c_3 only depending on c such that, when $m \leq c_3\left(\frac{\log n}{p^2\phi(\gamma)} \vee \frac{\log(1/(p^2\gamma))}{p^2\gamma}\right)$, for any estimator $\hat{\pi}$

$$\mathbb{P}\left[\operatorname{overlap}(\pi^*, \hat{\pi}) < 1\right] \geq 1 - c,$$

where π^* is uniformly distributed over $\mathcal{S}_{n,m}$.

Theorem 5 implies, for the purpose of exact recovery, the threshold for the number of correlated nodes m is of the order $\frac{\log n}{p^2\phi(\gamma)} \vee \frac{\log(1/(p^2\gamma))}{p^2\gamma}$. Under the weak signal regime $\gamma = O(1)$, we obtain the same rate as for partial recovery described in Theorem 4. Although the $\log n$ scaling has been observed in many other problems on random graphs, under the strong signal regime $\gamma = \omega(1)$, Theorem 5 highlights a transition from $\frac{\log n}{p^2\phi(\gamma)}$ to $\frac{\log(1/(p^2\gamma))}{p^2\gamma}$ if $\log^2\frac{1}{p} - \log^2\frac{1}{\rho} \gtrsim \log n$. In the latter regime, the difficulty is essentially the recovery of mapping given the sets of correlated nodes (S^*, T^*). See more discussions in Section 4.

In comparison to prior work, our results of partial recovery in Theorem 4 match the thresholds established in Wu et al. (2022) up to a constant factor in both dense and sparse regimes for the special case $S^* = V(G_1)$. Furthermore, the threshold $\frac{\log(1/(p^2\gamma))}{p^2\gamma}$ for exact recovery is derived from addressing the alignment problem for the subgraphs with the additional information on the domain and range of π^*, which applies the result in Wu et al. (2022).

1.2. Related Work

Graph sampling. Graph sampling methodologies are often propelled by many practical factors. Most notably, these encompass data scarcity, high data acquisition costs (Stumpf et al., 2005), and limited surveys of hidden structures (Lancichinetti and Fortunato, 2009; Yang et al., 2013; Fortunato and Hric, 2016). In scenarios where observations are sampled from two large networks, it becomes unrealistic to presume that correlation exists among all nodes within the sampled subgraphs. As a result, a pair of partially correlated graphs emerge naturally. While the precise number of correlated nodes may not be accessible, we often have some partial knowledge on the scale. For instance, when the observations are induced subgraphs of randomly selected nodes, the number of correlated nodes follows a hypergeometric distribution that concentrates around the mean value.

Besides the recent literature on the graph alignment problem, the correlation detection is another related topic. Given a pair of graphs, their correlation detection is formulated as a hypothesis testing problem, wherein the null hypothesis assumes independent random graphs, while the alternative assumes edge correlation under a latent permutation. Barak et al. (2019) proposed a hypothesis testing model for correlated Erdős-Rényi graphs and provided a pseudo-polynomial time algorithm for detection under certain conditions on the edge connection probability and average degree. Wu et al. (2023) established the sharp threshold for dense Erdős-Rényi graphs and determined the threshold within a constant factor for sparse Erdős-Rényi graphs. Ding and Du (2023a) derived the sharp threshold for sparse Erdős-Rényi graphs by analyzing the densest subgraph. Additionally, Mao et al. (2021) proposed a polynomial time algorithm for detection by counting trees when the correlation coefficient exceeds a constant value. It is natural to ask whether the correlation can be detected when only a subsample from the graphs is collected. The probabilistic model is similar to the one present in the current paper, and we leave the exploration as our future work.

Efficient algorithms and computational hardness. Numerous algorithms have been developed for the recovery problem. For example, Yartseva and Grossglauser (2013) analyzed the percolation graph matching algorithm, Barak et al. (2019) analyzed the problem using subgraph matching techniques, and Mossel and Xu (2020) obtained an algorithm for the seeded setting based on a delicate analysis of local neighborhoods. However, these algorithms may be computationally inefficient. There are several polynomial-time algorithms for recovery, catering to different regimes correlation coefficients ρ. These include works by Babai et al. (1980); Bollobás (1982); Dai et al. (2019); Ganassali and Massoulié (2020); Ding et al. (2021); Mao et al. (2023a,c); Ding and Li (2023); Muratori and Semerjian (2024). For instance, Mao et al. (2023c) proposed a polynomial-time algorithm for recovery by counting chandeliers when the correlation coefficient $\rho > \sqrt{\alpha}$, where $\alpha \approx 0.338$ is the Otter's constant introduced in Otter (1948). Additionally, Ding and Li (2023) introduced an efficient iterative polynomial-time algorithm for sparse Erdős-Rényi graphs when the correlation coefficient is a constant.

It is postulated in (Hopkins and Steurer, 2017; Hopkins, 2018; Kunisky et al., 2019) that the framework of low-degree polynomial algorithms effectively demonstrates computation hardness of detecting and recovering latent structures, and it bears similarities to sum-of-square methods (Hopkins et al., 2017; Hopkins, 2018). Based on the conjecture on the hardness of low-degree polynomial algorithms, Mao et al. (2021) proved that there is no polynomial-time test or matching algorithm when the correlation coefficient satisfies $\rho^2 \leq \frac{1}{\text{polylog}(n)}$. Furthermore, Ding et al. (2023a) showed computation hardness for detection and exact recovery when $p = n^{-1+o(1)}$ and the correlation coefficient $\rho < \sqrt{\alpha}$, where $\alpha \approx 0.338$ is the Otter's constant, suggesting that several polynomial algorithms may be essentially optimal.

The maximal overlap in the form of (4) is a test statistic which aims to identify the mapping that maximizes the edge correlation between two graphs. It is known that finding the maximal overlap is an instance of quadratic assignment problem (QAP) (Pardalos et al., 1994), which is NP-hard to solve or to approximate (Makarychev et al., 2010). There are many studies aiming to detect or recover latent structures based on the maximal overlap statistics (Cullina and Kiyavash, 2016, 2017; Barak et al., 2019; Mossel and Xu, 2020; Ding et al., 2021; Wu et al., 2022, 2023; Hall and Massoulié, 2023). Finally, we mention that the recent work Ding et al. (2024) approximated the maximal overlap within a constant factor in polynomial-time for sparse Erdős-Rényi graphs, and Du et al. (2023) established a sharp transition on approximating problem on the performance of online algorithms for dense Erdős-Rényi graphs.

Other graph models. Many properties of the correlated Erdős-Rényi graphs model have been extensively investigated. However, the strong symmetry and tree-like structure inherent in this model distinguish it significantly from graph models encountered in practical applications. Therefore, it is crucial to explore more general graph models. One such model is inhomogeneous random graph model, where the edge connecting probability varies among edges in the graph (Rácz and Sridhar, 2023; Song et al., 2023; Ding et al., 2023b). Besides, geometric random graph model (Wang et al., 2022; Bangachev and Bresler, 2023; Sentenac et al., 2023; Gong and Li, 2024), planted cycle model (Mao et al., 2023b, 2024), planted subhypergraph model (Dhawan et al., 2023) and corrupt model (Ameen and Hajek, 2023) have also been subjects of recent studies.

1.3. Notations

For any $n \in \mathbb{N}$, let $[n] \triangleq \{1, 2, ..., n\}$. For any $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max(a, b)$. We use standard asymptotic notation: for two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$, if $a_n \le Cb_n$ for some absolute constant C and for all n; $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$, if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ or $a_n \asymp b_n$, if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$; $a_n = o(b_n)$ or $b_n = \omega(a_n)$, if $a_n/b_n \to 0$ as $n \to \infty$.

For a given graph G, let $V(G)$ denote its vertex set and $E(G)$ denote its edge set. Let $\mathsf{v}(G) = |V(G)|$ denote the order of G and $\mathsf{e}(G) = |E(G)|$ denote size of G. For a set V, let $\binom{V}{2} \triangleq \{\{x, y\} : x, y \in V, x \neq y\}$ denote the collection of all subsets of V of cardinality two. We also write uv to denote an edge $\{u, v\}$. The induce subgraph of G over a vertex set V is denoted by $G[V]$. Given an injective mapping of vertices $\pi : S \subseteq V(G_1) \mapsto V(G_2)$, the induced injective mapping of edges is defined as $\pi^{\mathsf{E}} : \binom{S}{2} \mapsto \binom{V(G_2)}{2}$ as $\pi^{\mathsf{E}}(uv) = \pi(u)\pi(v)$ for $u, v \in S$. We also succinctly write $\pi(e) = \pi^{\mathsf{E}}(e)$ for an edge e when the meaning is clear from the context.

2. Correlated functional digraph

A mapping from a set to itself can be graphically represented as *functional digraph* (see, e.g., (West, 2021, Definition 1.3.3)). Here we extend the notion to a mapping with different domain and range sets, where the elements from the two sets are correlated. While our focus in this section is on the mapping between the edges in G_1 and G_2, the graphical representation can be easily extended to mappings between two arbitrary finite sets such as vertices.

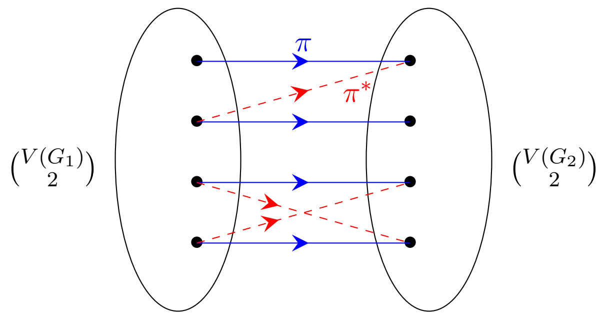

Figure 1: Examples of the mapping π and the underlying correlation π^*, where the domain and range of π and π^* could be different.

We first provide an equivalent description of the estimator in (4). Given a domain subset $S \subseteq \binom{V(G_1)}{2}$ and an injective function $\pi : S \mapsto \binom{V(G_2)}{2}$, we define the *intersection graph* $\mathcal{H}_\pi$ as

$$V(\mathcal{H}_\pi) = V(G_1), \quad e \in E(\mathcal{H}_\pi) \text{ if and only if } e \in E(G_1) \cap S \text{ and } \pi(e) \in E(G_2).$$

The estimator (4) maximizes the size of the intersection graph $|E(\mathcal{H}_\pi)|$. More generally, in our analysis in Section 3, we need to count the number of edges present in some subset $\mathcal{E} \subseteq S$ given by

$$|\mathcal{E} \cap E(\mathcal{H}_\pi)| = \sum_{e \in \mathcal{E}} \mathbf{1}_{\{e \in E(\mathcal{H}_\pi)\}} = \sum_{e \in \mathcal{E}} \mathbf{1}_{\{e \in E(G_1)\}} \mathbf{1}_{\{\pi(e) \in E(G_2)\}}. \tag{5}$$

Due to the correlation between the edges in G_1 and G_2, the counters $\mathbf{1}_{\{e \in E(\mathcal{H}_\pi)\}}$ are correlated random variables. The main idea is to decompose $\mathcal{E}$ into independent parts. Specially, the correlation is prescribed by the underlying mapping π^* as illustrated in Figure 1, where the correlated edges are red dashed lines. To formally describe all correlation relationships, we introduce the *correlated functional digraph* of a mapping π between a pair of graphs.

Definition 6 (Correlated functional digraph) *Let $\pi^* : S^* \mapsto T^*$ be the underlying mapping between correlated elements. The correlated functional digraph of the function $\pi : S \mapsto T$ is constructed as follows. Let the vertex set be $S \cup S^* \cup T \cup T^*$. We first add every edge $e \mapsto \pi(e)$ for $e \in S$, and then merge each pair of nodes $(e, \pi^*(e))$ for $e \in S^*$ into one node.*

It should be noted that both π and π^* are injective mappings under our model. After merging all pairs of nodes under π^*, the degree of each vertex in the correlated functional digraph is at most two. Therefore, the connected components consist of paths and cycles, where the self-loop is understood as a cycle of length one. The connected components are illustrated in Figure 2. Let $\mathcal{P}$ and $\mathcal{C}$ denote the collections of subsets of $\mathcal{E}$ belonging to different connected paths and cycles, respectively. Note that the sets from $\mathcal{P}$ and $\mathcal{C}$ are disjoint. Consequently,

$$|\mathcal{E} \cap E(\mathcal{H}_\pi)| = \sum_{P \in \mathcal{P}} |P \cap E(\mathcal{H}_\pi)| + \sum_{C \in \mathcal{C}} |C \cap E(\mathcal{H}_\pi)|,$$

where the summands are mutually independent.

In our model, the edge correlations are assumed to be homogeneous, and hence the distribution of $|P \cap E(\mathcal{H}_\pi)|$ and $|C \cap E(\mathcal{H}_\pi)|$ only depends on the size of the component. Let $\kappa^{\mathsf{P}}_\ell(t)$ and $\kappa^{\mathsf{C}}_\ell(t)$ denote the cumulant generating functions of $|P \cap E(\mathcal{H}_\pi)|$ and $|C \cap E(\mathcal{H}_\pi)|$ with $|P| = |C| = \ell$, respectively, and we have

$$\log \mathbb{E}\left[e^{t|P\cap E(\mathcal{H}_\pi)|}\right] = \kappa^{\mathsf{P}}_{|P|}(t), \qquad \log \mathbb{E}\left[e^{t|C\cap E(\mathcal{H}_\pi)|}\right] = \kappa^{\mathsf{C}}_{|C|}(t).$$

The lower-order cumulants can be promptly calculated. For instance,

$$\kappa^{\mathsf{C}}_1(t) = \log(1 + p_{11}(e^t - 1)), \tag{6}$$

$$\kappa^{\mathsf{C}}_2(t) = \log(1 + 2p^2(e^t - 1) + p_{11}^2(e^t - 1)^2). \tag{7}$$

It is however essential to establish upper bounds for higher-order cumulants in terms of lower-order ones. To this end, we introduce the following lemma.

Lemma 7 *For any $\rho > 0$, $0 < p < 1$, and $t > 0$,*

$$\kappa^{\mathsf{P}}_1(t) \le \frac{1}{2}\kappa^{\mathsf{C}}_2(t) \le \kappa^{\mathsf{C}}_1(t) \quad \text{and} \quad \kappa^{\mathsf{P}}_\ell(t) \le \kappa^{\mathsf{C}}_\ell(t) \le \frac{\ell}{2}\kappa^{\mathsf{C}}_2(t), \quad \forall\, \ell \ge 2.$$

Consequently,

$$\log \mathbb{E}\left[e^{t|\mathcal{E}\cap E(\mathcal{H}_\pi)|}\right] \le \frac{|\mathcal{E}|}{2}\kappa^{\mathsf{C}}_2(t) + L\left(\kappa^{\mathsf{C}}_1(t) - \frac{1}{2}\kappa^{\mathsf{C}}_2(t)\right), \tag{8}$$

where L denotes the number of self-loops.

The proof of Lemma 7 is deferred to Section C.1. The special case that both π and π^* are bijective has been studied in Wu et al. (2022); Ding and Du (2023b); Hall and Massoulié (2023), the correlation relationships under which can be characterized by a permutation $(\pi^*)^{-1} \circ \pi$. In this case, the connected components of the functional digraph of permutations are all cycles. However, in our case, the domain and range of π and π^* could be different and we need to deal with delicate correlations among the edges involving both cycles and paths by Lemma 7.

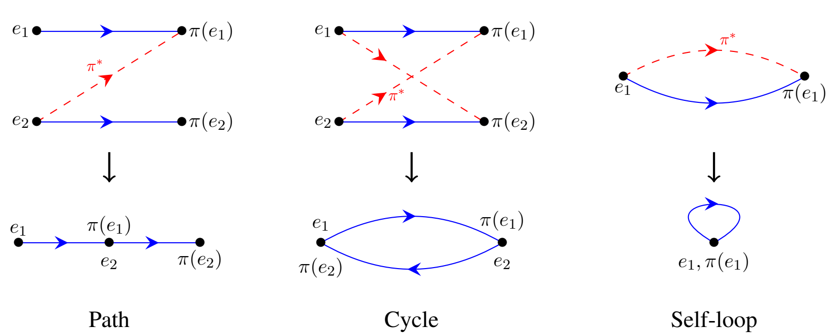

Figure 2: The connected components in the correlated functional digraph.

3. Recovery by maximizing the size of intersection graph

In this section, we prove the possibility results by analyzing the estimator $\hat{\pi}$ given in (4). By the optimality condition, it suffices to show that, for any $\pi^* \in \mathcal{S}_{n,m}$, we have $e(\mathcal{H}_{\pi^*})$ exceeds $\max_{\pi:d(\pi,\pi^*)>\tau} e(\mathcal{H}_\pi)$ with high probability when the underlying correlation is specified by π^*, where the thresholds $\tau = 0$ and δm are for exact and partial recoveries, respectively. In the following, we fix π^* and provide a general recipe for the upper bound of $\mathbb{P}_{\pi^*}[d(\hat{\pi}, \pi^*) = k]$. The overall error probability follows from the summation over the desired range of k.

Let $\mathcal{T}_k \subseteq \mathcal{S}_{n,m}$ denote the set of injections π such that $d(\pi, \pi^*) = k$. For $\pi \in \mathcal{T}_k$, by definition, there exists a set of correctly matched vertices (the self-loops in the correlated functional digraph of π over the vertices), denoted by $F_\pi \triangleq \{v \in S^* \cap S : \pi^*(v) = \pi(v)\}$ of cardinality $|F_\pi| = m - k$. The induced subgraphs of $\mathcal{H}_\pi$ and $\mathcal{H}_{\pi^*}$ over F_π are identical. Therefore,

$$e(\mathcal{H}_\pi) \geq e(\mathcal{H}_{\pi^*}) \iff e(\mathcal{H}_\pi) - e(\mathcal{H}_\pi[F_\pi]) \geq e(\mathcal{H}_{\pi^*}) - e(\mathcal{H}_{\pi^*}[F_\pi]).$$

It should be noted that correlated random variables are contained within the two sides of the inequality. Nevertheless, for any threshold τ_k, either $\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k$ or $\mathsf{e}(\mathcal{H}_\pi) - \mathsf{e}(\mathcal{H}_\pi[F_\pi]) \geq \tau_k$ holds. Therefore, we have the following upper bound:

$$\{d(\hat{\pi}, \pi^*) = k\} \subseteq \bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k\} \cup \{\mathsf{e}(\mathcal{H}_\pi) - \mathsf{e}(\mathcal{H}_\pi[F_\pi]) \geq \tau_k\}.$$

The first event is indicative of a weak signal, while the latter implies the presence of strong noise. The crucial result to establish is that, for a suitable threshold τ_k, both bad events will occur with a low probability. Here we may pick τ_k a function of all other parameters m, k, p, ρ. For brevity we also write $\tau_k = \tau(m, k, p, \rho)$.

Bad event of signal. For a fixed $\pi \in \mathcal{T}_k$, the random variable $\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi])$ counts the total number of edges among $N_k \triangleq \binom{m}{2} - \binom{m-k}{2} = mk(1 - \frac{k+1}{2m})$ pairs of vertices, where each edge presents independently with probability p_{11}. Furthermore, F_π is a subset of S^* of cardinality $m-k$. While the size of $\mathcal{T}_k$ could be large, the total number of possible F_π is at most $\binom{m}{m-k} = \binom{m}{k}$. Therefore,

$$\begin{aligned}\mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k\}\right] &\leq \mathbb{P}\left[\bigcup_{\substack{F \subseteq S^* \\ |F| = m-k}} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F]) < \tau_k\}\right] \\ &\leq \binom{m}{k} \mathbb{P}\left[\mathrm{Bin}(N_k, p_{11}) < \tau_k\right]. \end{aligned} \tag{9}$$

For $\tau_k < N_k p_{11}$, the tail of binomial distributions follows from the standard Chernoff bound.

Bad event of noise. The analyses for the noise part is more involved due to the mismatch between π and the underlying π^*. Let S_π denote the domain of π, and $\mathcal{E}_\pi \triangleq \binom{S_\pi}{2} - \binom{F_\pi}{2}$. Then the total number of edges $\mathsf{e}(\mathcal{H}_\pi) - \mathsf{e}(\mathcal{H}_\pi[F_\pi])$ can be equivalently represented as $|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)|$, and the cumulant generating function has been upper bounded in Lemma 7 thanks to the decomposition based on the correlated functional digraph. Thus, the error probability can be obtained via the Chernoff bound by optimizing over $t > 0$ in (8).

To this end, we need to upper bound the number of self-loops in (8). For a self-loop over an edge $e = uv$, we have $\pi(uv) = \pi^*(uv)$. Note that $\mathcal{E}_\pi$ excludes the edges in the induced subgraph over F_π. It necessarily holds that $\pi(u) = \pi^*(v)$ and $\pi(v) = \pi^*(u)$, which contributes two mismatched vertices in the reconstruction of the underlying mapping. Since the total number of mismatched vertices for $\pi \in \mathcal{T}_k$ equals to k, the number of self-loops is at most $\frac{k}{2}$. Consequently, applying (8) with the formula of lower-order cumulants (6) and (7) yields the following lemma, whose proof is deferred to Section C.2.

Lemma 8 *If $\tau_k > |\mathcal{E}_\pi| p^2$, then*

$$\mathbb{P}\left[|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)| \geq \tau_k\right] \leq \exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right) + \frac{\tau_k}{2} - \frac{|\mathcal{E}_\pi|p^2}{2} + \frac{k\gamma}{4(2+\gamma)}\right).$$

In view of Lemma 8, for $\tau_k > |\mathcal{E}_\pi| p^2$, we can apply the union bound for the probability of the bad event due to noise. It remains to upper bound the cardinality of $\mathcal{T}_k$. We first choose $m - k$ elements from the domain of π^* and map them to the same value as π^*. Then, the remaining domain and range of size k and the mapping are selected arbitrarily. Then we obtain

$$|\mathcal{T}_k| \leq \binom{m}{m-k}\binom{n-m+k}{k}^2 k! \leq \frac{m^k n^{2k}}{k!^2},$$

where the last step applies the upper bound $\binom{n}{k} \leq \frac{n^k}{k!}$. Since $e^{\frac{k\gamma}{4(2+\gamma)}} \leq e^{k/4}$, we have

$$\begin{aligned}\mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_\pi) - \mathsf{e}(\mathcal{H}_\pi[F_\pi]) \geq \tau_k\}\right] &= \mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k}\{|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)| \geq \tau_k\}\right] \\ &\leq |\mathcal{T}_k| e^{\frac{k\gamma}{4(2+\gamma)}} \exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right) + \frac{\tau_k}{2} - \frac{|\mathcal{E}_\pi|p^2}{2}\right) \\ &\leq n^{3k}\exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right) + \frac{\tau_k}{2} - \frac{|\mathcal{E}_\pi|p^2}{2}\right). \end{aligned} \tag{10}$$

The following propositions provide sufficient conditions on m for partial and exact recoveries.

Proposition 9 (Upper bound for partial recovery) *For any $\delta \in (0,1)$, there exists a constant $c_1(\delta) > 0$ such that, when $m \geq \frac{c_1(\delta)\log n}{p^2\phi(\gamma)}$, for any $\pi^* \in \mathcal{S}_{n,m}$, the estimator in (4) satisfies*

$$\mathbb{P}\left[\mathrm{overlap}(\hat{\pi}, \pi^*) < \delta\right] \leq (\log n)^{-1+o(1)}.$$

Proposition 10 (Upper bound for exact recovery) *There exists a universal constant $C > 0$ such that, when $m \geq C\left(\frac{\log(1/(p^2\gamma))}{p^2\gamma} \vee \frac{\log n}{p^2\phi(\gamma)}\right)$, for any $\pi^* \in \mathcal{S}_{n,m}$, the estimator in (4) satisfies*

$$\mathbb{P}\left[\hat{\pi} \neq \pi^*\right] \leq \frac{\exp(-\log m)}{1 - \exp(-\log m)} + \frac{\exp(-\log n)}{1 - \exp(-\log n)}.$$

By Propositions 9 and 10, we prove the possibility results in Theorems 4 and 5. The proofs of Propositions 9 and 10 are deferred to Sections A and B, respectively.

4. Impossibility results

In this section, we present the impossibility results for the graph alignment problem. Under our proposed model, the alignment problem aims to recover the domain $S^* \subseteq V(G_1)$, range $T^* \subseteq V(G_2)$, and the mapping $\pi^* : S^* \mapsto T^*$. When equipped with the additional knowledge on S^* and T^*, our problem can be reduced to recovery with full observations on smaller graphs, the reconstruction threshold for which is settled in Wu et al. (2022). The lower bound therein remains valid when the number of correlated nodes is substituted with m. However, such reduction only proves tight in a limited number of regimes (see Proposition 12). We will establish the impossibility results for the remaining regimes by Fano's method. Two main ingredients of Fano's method are outlined as follows:

- Construct a packing set $\mathcal{M}$ of the parameter space $\mathcal{S}_{n,m}$ such that any two distinct elements from $\mathcal{M}$ differ by a prescribed threshold. Specifically, in partial recovery, the overlap of each pair is less than δ, which is equivalent to $\min_{\pi \neq \pi' \in \mathcal{M}} d(\pi, \pi') \geq (1-\delta)m$, while in exact recovery $\mathcal{M} = \mathcal{S}_{n,m}$. The cardinality of $\mathcal{M}$ measures the complexity of the parameter space under the target metric.

- Choose the uniform prior on π^* over $\mathcal{M}$ and upper bound the mutual information $I(\pi^*; G_1, G_2)$. Given π^*, the conditional distribution of the observed graphs (G_1, G_2) is specified in Definition 3. For the mutual information, let $\mathcal{P}$ denote the joint distribution of (G_1, G_2) and $\mathcal{Q}$ be any distribution over (G_1, G_2), then

$$I(\pi^*; G_1, G_2) = \mathbb{E}_{\pi^*}[D(\mathcal{P}_{G_1,G_2|\pi^*} \| \mathcal{P}_{G_1,G_2})] \leq \max_{\pi} D(\mathcal{P}_{G_1,G_2|\pi} \| \mathcal{Q}_{G_1,G_2}). \tag{11}$$

  The impossibility results follows if $I(\pi^*; G_1, G_2) \leq c \log |\mathcal{M}|$ for some small constant c.

Let $\mathcal{M}_\delta$ denote a packing set under the overlap threshold δ. The size of $\mathcal{M}_\delta$ follows from the standard volume argument (Polyanskiy and Wu, 2022, Theorem 27.3). For $r \in [m]$, let $B(\pi, r) \triangleq \{\pi' : d(\pi, \pi') \leq r\}$ denote the ball of radius r centered at π. Then we have

$$|\mathcal{M}_\delta| \geq \frac{|\mathcal{S}_{n,m}|}{\max_\pi |B(\pi, (1-\delta)m - 1)|} \geq \frac{|\mathcal{S}_{n,m}|}{\max_\pi |B(\pi, (1-\delta)m)|}.$$

It remains to evaluate the cardinality of $\mathcal{S}_{n,m}$ and upper bound the volume of the ball under our distance metric d. It is straightforward to obtain that $|\mathcal{S}_{n,m}| = \binom{n}{m}^2 m!$. Let $k = \delta m$. Note that all elements from $B(\pi, m-k)$ have at least k common mappings. To upper bound $|B(\pi, m-k)|$, we first choose k elements from the domain of π and map to the same value as π, and the remaining domain and range of size $m-k$ and the mapping are selected arbitrarily. We get $|B(\pi, m-k)| \leq \binom{m}{k}\binom{n-k}{m-k}^2 (m-k)!$. Consequently,

$$|\mathcal{M}_\delta| \geq \frac{\binom{n}{m}^2 m!}{\binom{m}{k}\binom{n-k}{m-k}^2 (m-k)!} = \left(\frac{\binom{n}{k}}{\binom{m}{k}}\right)^2 k! > \left(\frac{n^2 k}{e^3 m^2}\right)^k \geq \left(\frac{\delta n}{e^3}\right)^k, \tag{12}$$

where we use the inequalities that $(\frac{n}{k})^k \leq \binom{n}{k} < (\frac{en}{k})^k$ and $k! \geq (k/e)^k$. Fano's method provides a lower bound on the Bayesian risk when π is uniformly distributed over $\mathcal{M}_\delta$, which further lower bound the minimax risk. The above argument also yields a lower bound when π is uniform over

$\mathcal{S}_{n,m}$ via generalized Fano's inequality (Banerjee et al., 2012, Lemma 20). The following propositions provide lower bounds for m for partial recovery and exact recovery, and thus prove the lower bounds in Theorems 4 and 5.

Proposition 11 (Lower bound for partial recovery) *For any $\delta \in (0,1)$, if $m \le \frac{c \log n}{p^2 \phi(\gamma)}$, then for any estimator $\hat{\pi}$,*

$$\mathbb{P}\left[\mathrm{overlap}(\hat{\pi}, \pi^*) < \delta\right] \ge 1 - \frac{13c}{\delta}.$$

Proof For any π with domain S and range T such that $|S| = |T| = m$, arbitrarily pick a bijection $\sigma : V(G_1) \mapsto V(G_2)$ such that $\sigma|_S = \pi$. Then, the conditional distribution $\mathcal{P}_{G_1,G_2|\pi}$ can be factorized into

$$\mathcal{P}_{G_1,G_2|\pi} = \prod_{e \in \binom{S}{2}} P(e, \pi(e)) \prod_{e \in \binom{V(G_1)}{2} \setminus \binom{S}{2}} Q(e, \sigma(e)),$$

where $P \sim \mathrm{Bern}(p, p, \rho)$ and $Q \sim \mathrm{Bern}(p, p, 0)$. Pick $\mathcal{Q}$ in (11) to be an auxiliary null model under which G_1 and G_2 are independent with the same marginal as $\mathcal{P}$. Then, $\mathcal{Q}_{G_1,G_2}$ can be factorized into

$$\mathcal{Q}_{G_1,G_2} = \prod_{e \in \binom{S}{2}} Q(e, \pi(e)) \prod_{e \in \binom{V(G_1)}{2} \setminus \binom{S}{2}} Q(e, \sigma(e)).$$

The KL-divergence between the product measures $\mathcal{P}_{G_1,G_2|\pi}$ and $\mathcal{Q}_{G_1,G_2}$ can be expressed as

$$D\left(\mathcal{P}_{G_1,G_2|\pi} \| \mathcal{Q}_{G_1,G_2}\right) = \tbinom{m}{2} D(P\|Q)$$

for any $\pi : S \mapsto T$ with $|S| = |T| = m$. Applying Lemma 16, we obtain

$$\max_{\pi} D(\mathcal{P}_{G_1,G_2|\pi} \| \mathcal{Q}_{G_1,G_2}) \le \binom{m}{2} D(P\|Q) \le 25\tbinom{m}{2} p^2 \phi(\gamma). \tag{13}$$

Applying generalized Fano's inequality (Banerjee et al., 2012, Lemma 20) with (12) and (13), we obtain

$$\mathbb{P}\left[\mathrm{overlap}(\pi^*, \hat{\pi}) < \delta\right] \ge 1 - \frac{25\binom{m}{2} p^2 \phi(\gamma)}{\delta m \log\left(\frac{\delta n}{e^3}\right)} \ge 1 - \frac{13c}{\delta},$$

where π^* is uniformly distributed over $\mathcal{S}_{n,m}$. ∎

Proposition 12 (Lower bound for exact recovery) *For any $c \in (0,1)$ and any estimator $\hat{\pi}$, there exists constant c_3 only depending on c such that, when $m \le c_3 \left(\frac{\log n}{p^2 \phi(\gamma)} \vee \frac{1}{p^2 \gamma} \log\left(\frac{1}{p^2 \gamma}\right)\right)$,*

$$\mathbb{P}\left[\hat{\pi} \ne \pi^*\right] \ge 1 - c,$$

where π^* is uniformly distributed over $\mathcal{S}_{n,m}$.

Proof We first apply the reduction argument. With the additional information on the domain and range of π^*, our problem can be reduced to the reconstruction of mapping as in Wu et al. (2022). Applying the lower bound in (Wu et al., 2022, Theorem 4), for a fixed $\epsilon \in (0, 1)$, when

$m(\sqrt{p_{00}p_{11}} - \sqrt{p_{01}p_{10}})^2 \le (1-\epsilon)\log m$, we have $\mathbb{P}[\hat{\pi} \ne \pi^*] \ge 1 - o(1)$ for any estimator $\hat{\pi}$. Note that $(\sqrt{p_{00}p_{11}} - \sqrt{p_{01}p_{10}})^2 \asymp p^2(\gamma \wedge \gamma^2) \asymp (\rho^2) \wedge (\rho p)$. Therefore, when

$$m \lesssim \frac{1}{p^2(\gamma \wedge \gamma^2)} \log\left(\frac{1}{p^2(\gamma \wedge \gamma^2)}\right), \tag{14}$$

we have $\mathbb{P}[\hat{\pi} \ne \pi^*] \ge 1 - o(1)$. Applying Proposition 11 with $\delta = 1/2$ yields that, when

$$m \lesssim \frac{\log n}{p^2 \phi(\gamma)}, \tag{15}$$

we have $\mathbb{P}[\hat{\pi} \ne \pi^*] \ge 1 - c$ for $c \in (0,1)$.

When $\frac{1}{p^2(\gamma \wedge \gamma^2)} \asymp n$, by (14), exact recovery is impossible, even when $m = n$. Next we consider the regime that $\frac{1}{p^2(\gamma \wedge \gamma^2)} \lesssim n$. When $\gamma \le 1$, we have $p^2(\gamma \wedge \gamma^2) = p^2\gamma^2 \asymp p^2\phi(\gamma)$, and thus

$$\frac{1}{p^2(\gamma \wedge \gamma^2)} \log\left(\frac{1}{p^2(\gamma \wedge \gamma^2)}\right) \lesssim \frac{\log n}{p^2\phi(\gamma)}.$$

When $\gamma \ge 1$, $\gamma \wedge \gamma^2 = \gamma$. By comparing (14) and (15), we derive that exact recovery is impossible if $m \lesssim \frac{\log n}{p^2\phi(\gamma)} \vee \frac{1}{p^2\gamma}\log\left(\frac{1}{p^2\gamma}\right)$. ■

5. Discussion and future directions

This paper proposes the partially correlated Erdős-Rényi graphs model, wherein a pair of induced subgraphs with a certain size are correlated. We investigate the optimal information-theoretic threshold for recovering the latent correlated subgraphs and the hidden vertices correspondence under our new model. In comparison with prior work on correlated Erdős-Rényi graphs model, the additional challenge arises from the unknown location of the correlated subsets. For a candidate mapping π whose domain may include both correlated and ambient subgraphs, we extend the classical notion of functional digraph to formally describe the correlation structure among the edges. We observe from the correlated functional digraph that the independent components consist of cycles and paths. The graphical representation may be of independent interest for general models.

There are many problems to be further investigated under our proposed model:

- *Refined results.* The results in the paper could be further refined in various ways, such as deriving the sharp constants and characterizing the optimal scaling in terms of the fraction δ in partial recovery.
- *Efficient algorithms.* It is of interest to investigate the polynomial-time algorithms and identify the computational hardness under our model. More efficient algorithms are also desirable when the signal is stronger.
- *Graph sampling.* One motivation of the paper stems from graph sampling as discussed in Section 1.2. The sampled subgraphs are partially correlated, where the size of correlated subsets is a random variable depending on the sampling methods. Thus, it is natural to ask about the sample size needed for reliable recovery.
- *Correlation test.* The correlation test problem under our model is also highly relevant. It is interesting to find out whether the detection problem is strictly easier than recovery, both in terms of the information thresholds and algorithmic developments.

Acknowledgments

The work of P. Yang was supported by the grant of National Natural Science Foundation of China (NSFC) 12101353.

References

Taha Ameen and Bruce Hajek. Robust graph matching when nodes are corrupt. *arXiv preprint arXiv:2310.18543*, 2023.

László Babai, Paul Erdős, and Stanley M Selkow. Random graph isomorphism. *SIaM Journal on computing*, 9(3):628–635, 1980.

Siddhartha Banerjee, Nidhi Hegde, and Laurent Massoulié. The price of privacy in untrusted recommendation engines. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 920–927. IEEE, 2012.

Kiril Bangachev and Guy Bresler. Detection of l_∞ geometry in random geometric graphs: Suboptimality of triangles and cluster expansion. *arXiv preprint arXiv:2310.14501*, 2023.

Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander C Berg, Tamara L Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondences. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 26–33. IEEE, 2005.

Béla Bollobás. Distinguishing vertices of random graphs. In *North-Holland Mathematics Studies*, volume 62, pages 33–49. Elsevier, 1982.

Timothee Cour, Praveen Srinivasan, and Jianbo Shi. Balanced Graph Matching. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

TM Cover and JA Thomas. Elements of information theory. hoboken, nj, usa: John wiely & sons, 2006.

Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for Erdős-Rényi graph matching. *ACM SIGMETRICS performance evaluation review*, 44(1):63–72, 2016.

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

Osman Emre Dai, Daniel Cullina, Negar Kiyavash, and Matthias Grossglauser. Analysis of a canonical labeling algorithm for the alignment of correlated Erdős-Rényi graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–25, 2019.

Abhishek Dhawan, Cheng Mao, and Alexander S Wein. Detection of dense subhypergraphs by low-degree polynomials. *arXiv preprint arXiv:2304.08135*, 2023.

Jian Ding and Hang Du. Detection threshold for correlated Erdős-Rényi graphs via densest subgraph. *IEEE Transactions on Information Theory*, 2023a.

Jian Ding and Hang Du. Matching recovery threshold for correlated random graphs. *The Annals of Statistics*, 51(4):1718–1743, 2023b.

Jian Ding and Zhangsong Li. A polynomial-time iterative algorithm for random graph matching with non-vanishing correlation. *arXiv preprint arXiv:2306.00266*, 2023.

Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179:29–115, 2021.

Jian Ding, Hang Du, and Zhangsong Li. Low-degree hardness of detection for correlated Erdős-Rényi graphs. *arXiv preprint arXiv:2311.15931*, 2023a.

Jian Ding, Yumou Fei, and Yuanzheng Wang. Efficiently matching random inhomogeneous graphs via degree profiles. *arXiv preprint arXiv:2310.10441*, 2023b.

Jian Ding, Hang Du, and Shuyang Gong. A polynomial-time approximation scheme for the maximal overlap of two independent Erdős-Rényi graphs. *Random Structures & Algorithms*, 2024.

Hang Du, Shuyang Gong, and Rundong Huang. The algorithmic phase transition of random graph alignment problem. *arXiv preprint arXiv:2307.06590*, 2023.

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

Luca Ganassali and Laurent Massoulié. From tree matching to sparse graph alignment. In *Conference on Learning Theory*, pages 1633–1665. PMLR, 2020.

Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Impossibility of partial recovery in the graph alignment problem. In *Conference on Learning Theory*, pages 2080–2102. PMLR, 2021.

Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Shuyang Gong and Zhangsong Li. The umeyama algorithm for matching correlated gaussian geometric models in the low-dimensional regime. *arXiv preprint arXiv:2402.15095*, 2024.

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

Georgina Hall and Laurent Massoulié. Partial recovery in the graph alignment problem. *Operations Research*, 71(1):259–272, 2023.

Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. In *International Colloquium on Automata, Languages, and Programming*, pages 594–604. Springer, 2010.

Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. Testing network correlation efficiently via counting trees. *arXiv preprint arXiv:2110.11816*, 2021.

Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Exact matching of random graphs with constant correlation. *Probability Theory and Related Fields*, 186(1-2):327–389, 2023a.

Cheng Mao, Alexander S Wein, and Shenduo Zhang. Detection-recovery gap for planted dense cycles. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2440–2481. PMLR, 2023b.

Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. Random graph matching at otter's threshold via counting chandeliers. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1345–1356, 2023c.

Cheng Mao, Alexander S. Wein, and Shenduo Zhang. Information-theoretic thresholds for planted dense cycles. *arXiv preprint arXiv:2402.00305*, 2024.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: an introduction to randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005. ISBN 978-0-521-83540-4.

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

Andrea Muratori and Guilhem Semerjian. Faster algorithms for the alignment of sparse correlated Erdős-Rényi random graphs. *arXiv preprint arXiv:2405.08421*, 2024.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pages 173–187. IEEE, 2009.

Richard Otter. The number of trees. *Annals of Mathematics*, pages 583–599, 1948.

Panos M Pardalos, Franz Rendl, and Henry Wolkowicz. The quadratic assignment problem: A survey and recent developments. 1994.

Erdős Paul and Rényi Alfréd. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6: 290–297, 1959.

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

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning. 2022.

Miklós Z Rácz and Anirudh Sridhar. Matching correlated inhomogeneous random graphs using the k-core estimator. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2499–2504. IEEE, 2023.

Flore Sentenac, Nathan Noiry, Matthieu Lerasle, Laurent Ménard, and Vianney Perchet. Online matching in geometric random graphs. *arXiv preprint arXiv:2306.07891*, 2023.

Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

Yukun Song, Carey E Priebe, and Minh Tang. Independence testing for inhomogeneous random graphs. *arXiv preprint arXiv:2304.09132*, 2023.

Michael PH Stumpf, Carsten Wiuf, and Robert M May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12):4221–4224, 2005.

Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Louis J Podrazik, Steven G Kratzer, Eric T Harley, Donniell E Fishkind, R Jacob Vogelstein, and Carey E Priebe. Fast approximate quadratic programming for large (brain) graph matching. *arXiv preprint arXiv:1112.5507*, 2011.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In *Conference on Learning Theory*, pages 3441–3488. PMLR, 2022.

Douglas B West. *Combinatorial mathematics*. Cambridge University Press, 2021.

Yihong Wu, Jiaming Xu, and Sophie H Yu. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory*, 68(8):5391–5417, 2022.

Yihong Wu, Jiaming Xu, and Sophie H Yu. Testing correlation of unlabeled random graphs. *The Annals of Applied Probability*, 33(4):2519–2558, 2023.

Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th international conference on data mining*, pages 1151–1156. IEEE, 2013.

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

Appendix A. Proof of Proposition 9

Let $\tau_k = N_k p_{11}(1-\eta)$ with

$$\eta = \sqrt{\frac{8h(\frac{k}{m})}{kp_{11}}} \cdot \mathbb{1}_{k \le m-1} + \sqrt{\frac{\log n}{kmp_{11}}} \cdot \mathbb{1}_{k=m},$$

where $h(x) \triangleq -x\log x - (1-x)\log(1-x)$ is the binary entropy function. Since $h(x)/x$ decreases in $(0,1)$, $h(x)/x \ge h(1-\delta)/(1-\delta)$ for $1-\delta \le x < 1$. By Lemma 17.5.1 in Cover and Thomas (2006), we have $\binom{m}{k} \le \exp[mh(k/m)]$ for any $k \le m-1$. By (9) and the Chernoff bound (21), when $k \le m-1$, we have

$$\mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k\}\right] \le \binom{m}{k} \exp\left(-\frac{N_k p_{11}\eta^2}{2}\right)$$
$$\le \exp\left(mh\left(\frac{k}{m}\right) - \frac{N_k p_{11}\eta^2}{2}\right) \le \exp\left(-mh\left(\frac{k}{m}\right)\right).$$

When $k = m$, since $N_k = \frac{mk}{2}\left(2 - \frac{k+1}{m}\right) \ge \frac{mk}{3}$, we have

$$\mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k\}\right] \le \binom{m}{k} \exp\left(-\frac{N_k p_{11}\eta^2}{2}\right) \le \exp\left(-\frac{\log n}{6}\right).$$

Pick $c_1(\delta) = 100 \vee \frac{200h(1-\delta)}{1-\delta}$. We then verify the condition in Lemma 14: $\eta \le \frac{\gamma}{4(1+\gamma)}$. Since $p^2\phi(\gamma) \le p^2\gamma(1+\gamma) \le 1$, we get $m \ge c_1(\delta)\log n$. Therefore,

$$\eta \le \sqrt{\frac{8h(1-\delta)}{(1-\delta)mp_{11}}} \cdot \mathbb{1}_{k \le m-1} + \sqrt{\frac{\log n}{m^2 p_{11}}} \cdot \mathbb{1}_{k=m} \le \left(\sqrt{\frac{8h(1-\delta)}{1-\delta}} \vee \frac{1}{\sqrt{c_1(\delta)}}\right) \frac{1}{\sqrt{mp_{11}}}$$
$$\le \left(\sqrt{\frac{8h(1-\delta)}{(1-\delta)c_1(\delta)}} \vee \frac{1}{c_1(\delta)}\right) \sqrt{\frac{\log(1+\gamma) - \gamma/(1+\gamma)}{\log n}} \le \frac{1}{5}\sqrt{\frac{\log(1+\gamma) - \gamma/(1+\gamma)}{\log n}}.$$

Recall the assumption stated in Section 1.1, where it's asserted that $p \ge n^{-1}$, thereby implying $\log(1+\gamma) \le \log n$. When $\gamma > 10$, $\eta \le \frac{1}{5} \le \frac{\gamma}{4(1+\gamma)}$. When $\gamma \le 10$, since $\log(1+x) - \frac{x}{1+x} - x^2 \le 0$ for any $x > 0$, $\sqrt{\frac{1}{mp_{11}}} \le \sqrt{\frac{\log(1+\gamma)-\gamma/(1+\gamma)}{\log n}} \le \frac{\gamma}{\sqrt{\log n}} \le \frac{\gamma}{4(1+\gamma)}$. Therefore, we obtain $\eta \le \frac{\gamma}{4(1+\gamma)}$. By Lemma 14, $\frac{\tau_k}{|\mathcal{E}_\pi|p^2} = (1+\gamma)(1-\eta) > 1$. Applying Lemma 8, we derive (10). Combining this with (19) in Lemma 14, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{\pi\in\mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_\pi)-\mathsf{e}(\mathcal{H}_\pi[F_\pi])\geq\tau_k\}\right]\\
\leq&n^{3k}\exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right)+\frac{\tau_k}{2}-\frac{|\mathcal{E}_\pi|p^2}{2}\right)\\
=&n^{3k}\exp\left\{-\frac{N_kp^2}{2}\phi\left[(1+\gamma)(1-\eta)-1\right]\right\}\leq n^{3k}\exp\left(-\frac{N_kp^2}{8}\phi(\gamma)\right).
\end{aligned}$$

Sum over $k\geq(1-\delta)m$, since $N_k\geq\frac{km}{3}$, we obtain

$$\begin{aligned}
&\sum_{k=\delta m}^{m}\mathbb{P}\left[d(\pi^*,\hat{\pi})=k\right]\\
\leq&\sum_{k=\delta m}^{m}\mathbb{P}\left[\bigcup_{\pi\in\mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_{\pi^*})-\mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi])<\tau_k\}\right]+\mathbb{P}\left[\bigcup_{\pi\in\mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_\pi)-\mathsf{e}(\mathcal{H}_\pi[F_\pi])\geq\tau_k\}\right]\\
\leq&\exp\left(-\frac{\log n}{6}\right)+\sum_{k=(1-\delta)m}^{m-1}\exp\left[-mh\left(\frac{k}{m}\right)\right]+\sum_{k=(1-\delta)m}^{m}\left[n^3\exp\left(-\frac{mp^2\phi(\gamma)}{24}\right)\right]^k\\
\leq&\exp\left(-\frac{\log n}{6}\right)+\frac{\exp\left[-(1-\delta)m\log n\right]}{1-\exp(-\log n)}+\sum_{k=(1-\delta)m}^{m-1}\exp\left[-mh\left(\frac{k}{m}\right)\right].
\end{aligned}$$

Combining this with Lemma 15, $\sum_{k=(1-\delta)m}^{m}\mathbb{P}\left[d(\pi^*,\hat{\pi})=k\right]\leq(\log n)^{-1+o(1)}$. We finish the proof.

Appendix B. Proof of Proposition 10

Let $\tau_k=N_kp_{11}(1-\eta)$ with $\eta=\frac{\gamma}{4(1+\gamma)}$, by (9) and the Chernoff bound (21),

$$\mathbb{P}\left[\bigcup_{\pi\in\mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_{\pi^*})-\mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi])<\tau_k\}\right]\leq\binom{m}{k}\exp\left(-\frac{N_kp_{11}\eta^2}{2}\right).$$

By Lemma 14, $\frac{\tau_k}{|\mathcal{E}_\pi|p^2}=(1+\gamma)(1-\eta)>1$. Applying Lemma 8, we derive (10). Combining this with (19) in Lemma 14, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{\pi\in\mathcal{T}_k}\{\mathsf{e}(\mathcal{H}_\pi)-\mathsf{e}(\mathcal{H}_\pi[F_\pi])\geq\tau_k\}\right]\\
\leq&n^{3k}\exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right)+\frac{\tau_k}{2}-\frac{|\mathcal{E}_\pi|p^2}{2}\right)\\
=&n^{3k}\exp\left\{-\frac{N_kp^2}{2}\phi\left[(1+\gamma)(1-\eta)-1\right]\right\}\leq n^{3k}\exp\left[-\frac{N_kp^2}{8}\phi(\gamma)\right].
\end{aligned}$$

Sum over $k \geq 1$, since $N_k \geq \frac{km}{3}$ and $\binom{m}{k} \leq m^k$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{m} \mathbb{P}\left[d(\pi^*, \hat{\pi}) = k\right] \\
&\leq \sum_{k=1}^{m} \mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi^*}) - \mathsf{e}(\mathcal{H}_{\pi^*}[F_\pi]) < \tau_k\}\right] + \mathbb{P}\left[\bigcup_{\pi \in \mathcal{T}_k} \{\mathsf{e}(\mathcal{H}_{\pi}) - \mathsf{e}(\mathcal{H}_{\pi}[F_\pi]) \geq \tau_k\}\right] \\
&\leq \sum_{k=1}^{m} \left[m \exp\left(-\frac{mp_{11}\eta^2}{6}\right)\right]^k + \left[n^3 \exp\left(-\frac{mp^2\phi(\gamma)}{24}\right)\right]^k.
\end{aligned}$$

Pick the universal constant $C = 400$. Recall that $\phi(\gamma) = (1+\gamma)\log(1+\gamma) - \gamma$. When $\gamma \leq 1$, since $\phi(\gamma) \leq \frac{\gamma^2}{1+\gamma}$, we obtain $p^2\phi(\gamma) \leq 16p_{11}\eta^2$. Therefore, $m \geq \frac{400 \log n}{p^2\phi(\gamma)}$ implies $m \geq \frac{12 \log m}{p_{11}\eta^2}$. When $\gamma > 1$, since $\gamma \leq 32(1+\gamma)\left[\frac{\gamma}{4(1+\gamma)}\right]^2$, we obtain $p^2\gamma \leq 32p_{11}\eta^2$. Since $m \geq \frac{400 \log(1/p^2\gamma)}{p^2\gamma}$, we have $m \geq \frac{384 \log m}{p^2\gamma} \geq \frac{12 \log m}{p_{11}\eta^2}$. Thus we get $m \exp\left(-\frac{mp_{11}\eta^2}{6}\right) \leq \exp(-\log m)$. When $m \geq \frac{400 \log n}{p^2\phi(\gamma)}$, we get $n^3 \exp\left(-\frac{mp^2\phi(\gamma)}{24}\right) \leq \exp(-\log n)$. Therefore, when $m \geq 400\left(\frac{\log(1/p^2\gamma)}{p^2\gamma} \vee \frac{\log n}{p^2\phi(\gamma)}\right)$, we have

$$m \exp\left(-\frac{mp_{11}\eta^2}{6}\right) \leq \exp(-\log m), n^3 \exp\left(-\frac{mp^2\phi(\gamma)}{24}\right) \leq \exp(-\log n).$$

Therefore, $\sum_{k=1}^{m} \mathbb{P}\left[d(\pi^*, \hat{\pi}) = k\right] \leq \frac{\exp(-\log m)}{1-\exp(-\log m)} + \frac{\exp(-\log n)}{1-\exp(-\log n)}$. We finish the proof.

Appendix C. Proof of Lemmas

C.1. Proof of Lemma 7

We first evaluate the moment generating function for paths. Consider a path P of size ℓ denoted by $\langle e_1 e_2 \dots e_\ell \rangle$ as illustrated in Figure 3. For each $i = 1, \dots, \ell$, define $A_{i-1} \triangleq \mathbf{1}_{\{e_i \in E(G_1)\}}$ and $B_i \triangleq \mathbf{1}_{\{\pi(e_i) \in E(G_2)\}}$. Then $(A_i, B_i) \sim \mathrm{Bern}(p, p, \rho)$. By definition (5),

$$|P \cap E(\mathcal{H}_\pi)| = \sum_{i=1}^{\ell} \mathbf{1}_{\{e_i \in E(G_1)\}} \mathbf{1}_{\{\pi(e_i) \in E(G_2)\}} = \sum_{i=1}^{\ell} A_{i-1}B_i.$$

For the sake of notational simplicity, we introduce an auxiliary random variable B_0 that is correlated with A_0 such that $(A_0, B_0) \sim \mathrm{Bern}(p, p, \rho)$. Then

$$\begin{aligned}
m_\ell \triangleq \mathbb{E}[e^{t|P \cap E(\mathcal{H}_\pi)|}] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{\ell} e^{tA_{i-1}B_i} | B_0 \dots B_\ell\right]\right] = \mathbb{E}\left[\prod_{i=1}^{\ell} \mathbb{E}\left[e^{tA_{i-1}B_i} | B_{i-1}B_i\right]\right] \\
&= \sum_{b_0, \dots, b_\ell \in \{0,1\}} \prod_{i=0}^{\ell} \mathbb{P}[B_i = b_i] \prod_{i=1}^{\ell} \mathbb{E}\left[e^{tA_{i-1}b_i} | B_{i-1} = b_{i-1}\right]. \qquad (16)
\end{aligned}$$

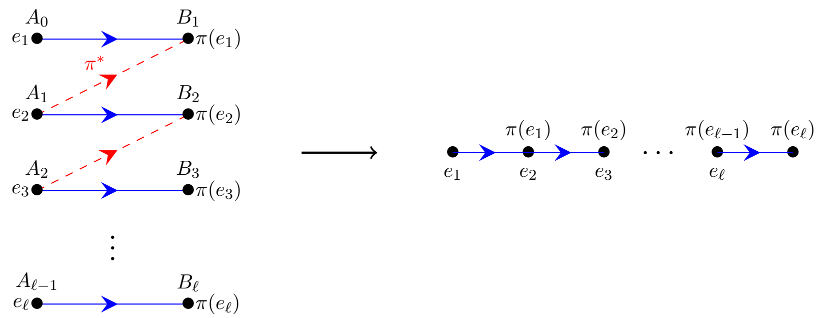

Figure 3: Illustration of a path of size ℓ.

Define $M(b_{i-1}, b_i) \triangleq \mathbb{P}[B_i = b_i]\mathbb{E}\left[e^{tA_{i-1}b_i}|B_{i-1} = b_{i-1}\right]$ for $b_{i-1}, b_i \in \{0, 1\}$ and a matrix

$$M \triangleq \begin{bmatrix} M(0,0) & M(0,1) \\ M(1,0) & M(1,1) \end{bmatrix} = \begin{bmatrix} \bar{p} & (\bar{p} + p_{01}(e^t - 1))p/\bar{p} \\ \bar{p} & p + p_{11}(e^t - 1) \end{bmatrix},$$

where $\bar{p} = 1 - p$. Recall that $\mathbb{P}[B_i = 1] = p$. Then we obtain that

$$m_\ell = \sum_{b_0,\dots,b_\ell \in \{0,1\}} \mathbb{P}[B_0 = b_0] M(b_0, b_1) \dots M(b_{\ell-1}, b_\ell) = [\bar{p}, p] M^\ell \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The trace and determinant of M is

$$T \triangleq \mathrm{Tr}(M) = 1 + p_{11}(e^t - 1), \quad D \triangleq \det(M) = \rho p \bar{p}(e^t - 1) > 0.$$

Since $D < p_{11}(e^t - 1)$, the discriminant is $T^2 - 4D > 0$. Hence, the matrix M has two distinct eigenvalues denoted by $\lambda_1 > \lambda_2 > 0$, and the general term of m_ℓ is

$$m_\ell = \alpha \lambda_1^\ell + \beta \lambda_2^\ell. \tag{17}$$

The coefficients α and β can be determined via the first two terms $m_0 = 1$ and m_1. Then we get

$$m_\ell = \left(\frac{1}{2} + \frac{2m_1 - T}{2\sqrt{T^2 - 4D}}\right)\lambda_1^\ell + \left(\frac{1}{2} - \frac{2m_1 - T}{2\sqrt{T^2 - 4D}}\right)\lambda_2^\ell.$$

Furthermore, by plugging $m_1 = 1 + p^2(e^t - 1)$, we get $T - m_1 = D$ and thus $m_1(T - m_1) > D$, which is equivalent to $|2m_1 - T| < \sqrt{T^2 - 4D}$. Therefore, both coefficients $\alpha, \beta \in (0, 1)$.

The analysis for cycles follows from similar arguments. Consider a cycle C of size ℓ denoted by $[e_1 \dots e_\ell]$ as illustrated in Figure 4. For each $i = 1, \dots, \ell$, define $A_{i-1} \triangleq \mathbf{1}_{\{e_i \in E(G_1)\}}$ and $B_i \triangleq \mathbf{1}_{\{\pi(e_i) \in E(G_2)\}}$. We also let $B_0 = B_\ell$ for notational simplicity. Then $(A_i, B_i) \sim \mathrm{Bern}(p, p, \rho)$ for $i = 0, \dots, \ell - 1$. Following a similar argument as (16), we have

$$\begin{aligned} \tilde{m}_\ell \triangleq \mathbb{E}[e^{t|C \cap E(\mathcal{H}_\pi)|}] &= \sum_{b_1,\dots,b_\ell = b_0 \in \{0,1\}} \prod_{i=1}^{\ell} \mathbb{P}[B_i = b_i] \prod_{i=1}^{\ell} \mathbb{E}\left[e^{tA_{i-1}b_i}|B_{i-1} = b_{i-1}\right] \\ &= \sum_{b_1,\dots,b_\ell = b_0 \in \{0,1\}} M(b_0, b_1) M(b_1, b_2) \dots M(b_{\ell-1}, b_0). \end{aligned}$$

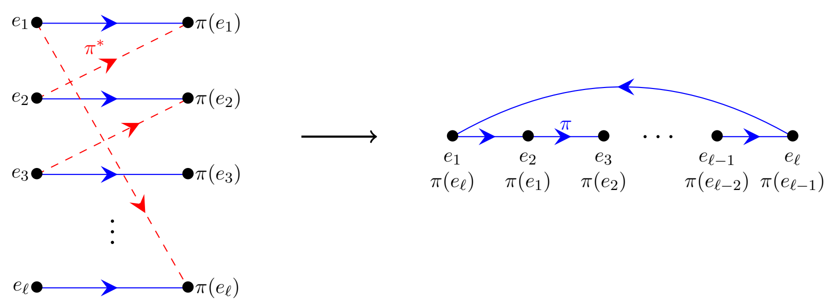

Figure 4: Illustration of a cycle of size ℓ.

Applying the eigenvalue decomposition of M again, we obtain that

$$\tilde{m}_\ell = \mathrm{Tr}(M^\ell) = \lambda_1^\ell + \lambda_2^\ell. \tag{18}$$

By definition, $\kappa_\ell^{\mathsf{P}}(t) = \log m_\ell$ and $\kappa_\ell^{\mathsf{C}}(t) = \log \tilde{m}_\ell$. To upper bound the cumulants, it suffices to consider m_ℓ and $\tilde{m}_\ell$. In (17), we have $\alpha, \beta \in (0,1)$ and $\lambda_1 > \lambda_2 > 0$. By monotonicity, it follows that $m_\ell \le \tilde{m}_\ell$ and thus

$$\kappa_\ell^{\mathsf{P}}(t) \le \kappa_\ell^{\mathsf{C}}(t).$$

For $x \in \mathbb{R}^n$ and $\ell \ge 2$, we have $\|x\|_\ell \le \|x\|_2 \le \|x\|_1$. It follows from the formula of $\tilde{m}_\ell$ in (18) that $\tilde{m}_\ell^{1/\ell} \le \tilde{m}_2^{1/2} \le \tilde{m}_1$. Equivalently,

$$\frac{1}{2}\kappa_2^{\mathsf{C}}(t) \le \kappa_1^{\mathsf{C}}(t), \quad \kappa_\ell^{\mathsf{C}}(t) \le \frac{\ell}{2}\kappa_2^{\mathsf{C}}(t) \quad \forall\, \ell \ge 2.$$

The last inequality $2\kappa_1^{\mathsf{P}}(t) \le \kappa_2^{\mathsf{C}}(t)$ follows by comparing the explicit formula $\kappa_1^{\mathsf{P}}(t) = \log(1 + p^2(e^t - 1))$ with $\kappa_2^{\mathsf{C}}(t)$ in (7) and using $p_{11} \ge p^2$.

Finally, since the summands over different connected components are independent, it follows that

$$\begin{aligned}
\log \mathbb{E}\left[e^{t|\mathcal{E}\cap E(\mathcal{H}_\pi)|}\right] &= \sum_{P\in\mathcal{P}} \kappa_{|P|}^{\mathsf{P}}(t) + \sum_{C\in\mathcal{C}} \kappa_{|C|}^{\mathsf{C}}(t) \\
&\le \sum_{P\in\mathcal{P}} \frac{|P|}{2}\kappa_2^{\mathsf{C}}(t) + \sum_{C\in\mathcal{C}:|C|\ge 2} \frac{|C|}{2}\kappa_2^{\mathsf{C}}(t) + \sum_{C\in\mathcal{C}:|C|=1} \kappa_1^{\mathsf{C}}(t) \\
&= \frac{|\mathcal{E}|}{2}\kappa_2^{\mathsf{C}}(t) + |\{C \in \mathcal{C} : |C| = 1\}| \left(\kappa_1^{\mathsf{C}}(t) - \frac{1}{2}\kappa_2^{\mathsf{C}}(t)\right),
\end{aligned}$$

where the last equality used fact that $|\mathcal{E}| = \sum_{P\in\mathcal{P}} |P| + \sum_{C\in\mathcal{C}} |C|$.

Remark 13 *We have two bounds for large ℓ in Lemma 7, namely $\kappa_\ell^{\mathsf{P}}(t) \le \kappa_\ell^{\mathsf{C}}(t)$ and $\kappa_\ell^{\mathsf{C}}(t) \le \frac{\ell}{2}\kappa_2^{\mathsf{C}}(t)$. For the first bound, we apply $\frac{1}{\ell}\log\left(\alpha\lambda_1^\ell + \beta\lambda_2^\ell\right) \le \frac{1}{\ell}\log(\lambda_1^\ell + \lambda_2^\ell)$, where $0 < \beta < \alpha < 1, \alpha + \beta = 1$ and $\lambda_1 > \lambda_2 > 0$. Consequently, $\lambda_1 - \frac{\log 2}{\ell} \le \frac{1}{\ell}\kappa_\ell^{\mathsf{P}}(t) \le \frac{1}{\ell}\kappa_\ell^{\mathsf{C}}(t) \le \lambda_1 + \frac{\log 2}{\ell}$. Hence, the first bound is essentially tight for large ℓ. The second bound, previously used in Wu et al. (2022), applies the inequality $\|x\|_\ell \le \|x\|_2$, which becomes less tight as ℓ increases. Nevertheless, it suffices for our analysis as the probability of long cycles occurring is relatively small.*

C.2. Proof of Lemma 8

By Lemma 7,

$$\log \mathbb{E}\left[e^{t|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)|}\right] \leq \frac{|\mathcal{E}_\pi|}{2}\kappa_2^{\mathsf{C}}(t) + L\left(\kappa_1^{\mathsf{C}}(t) - \frac{1}{2}\kappa_2^{\mathsf{C}}(t)\right),$$

where L denotes the number of self-loops. The self-loop for e only happens when $\pi(e) = \pi^*(e)$. For $uv \in \binom{V(G_1)}{2} \setminus \binom{F_\pi}{2}$, by the definition of F_π, $\pi(u) \neq \pi^*(u)$ or $\pi(v) \neq \pi^*(v)$. Therefore, $\pi(uv) = \pi^*(uv)$ implies that $\pi(u) = \pi^*(v)$ and $\pi(v) = \pi^*(u)$. Since $d(\pi^*, \pi) = k$, we must have $L \leq \frac{k}{2}$.

Applying the formulas (6) and (7) and the fact that $p_{11} \leq p$, we obtain

$$\begin{aligned}\kappa_2^{\mathsf{C}}(t) &\leq \log\left(1 + 2p^2(e^t - 1) + p^2(e^t-1)^2\right)\\ &= \log\left(1 + p^2(e^{2t}-1)\right) \leq p^2(e^{2t}-1)\end{aligned}$$

and

$$\begin{aligned}\kappa_1^{\mathsf{C}}(t) - \frac{1}{2}\kappa_2^{\mathsf{C}}(t) &= \frac{1}{2}\log\left[1 + \frac{2(p_{11}-p^2)}{p_{11}^2(e^t-1) + 2p^2 + (e^t-1)^{-1}}\right]\\ &\overset{(a)}{\leq} \frac{1}{2}\log\left[1 + \frac{2(p_{11}-p^2)}{2(p_{11}+p^2)}\right] \overset{(b)}{\leq} \frac{\gamma}{2(\gamma+2)},\end{aligned}$$

where (a) is because $x + x^{-1} \geq 2$ for any $x > 0$ and (b) is because $\log(1+x) \leq x$ for any $x \geq 0$. Therefore, we get

$$\log \mathbb{E}\left[e^{t|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)|}\right] \leq \frac{|\mathcal{E}_\pi|}{2}p^2(e^{2t}-1) + \frac{k\gamma}{4(\gamma+2)}.$$

For any $t > 0$, by the Chernoff bound,

$$\mathbb{P}[|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)| \geq \tau_k] \leq \exp\left(-t\tau_k + \frac{|\mathcal{E}_\pi|}{2}p^2(e^{2t}-1) + \frac{k\gamma}{4(2+\gamma)}\right).$$

Pick $t = \frac{1}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right)$. Then $t > 0$ by the assumption $\tau_k > |\mathcal{E}_\pi|p^2$. We obtain

$$\mathbb{P}\left[|\mathcal{E}_\pi \cap E(\mathcal{H}_\pi)| \geq \tau_k\right] \leq \exp\left(-\frac{\tau_k}{2}\log\left(\frac{\tau_k}{|\mathcal{E}_\pi|p^2}\right) + \frac{\tau_k}{2} - \frac{|\mathcal{E}_\pi|p^2}{2} + \frac{k\gamma}{4(2+\gamma)}\right).$$

C.3. Proof of Lemma 14

Lemma 14 *Recall that* $\phi(\gamma) = (1+\gamma)\log(1+\gamma) - \gamma$ *and* $\eta, \gamma > 0$. *If* $\eta \leq \frac{\gamma}{4(1+\gamma)}$, *then* $(1+\gamma)(1-\eta) > 1$ *and*

$$\phi\left[(1-\eta)(1+\gamma) - 1\right] \geq \frac{1}{4}\phi(\gamma). \tag{19}$$

Proof We note that $(1+\gamma)(1-\eta) \geq 1 + \gamma - \frac{\gamma}{4} > 1$ and

$$\begin{aligned}\phi\left[(1-\eta)(1+\gamma)-1\right] =&(1+\gamma)(1-\eta)\log\left[(1+\gamma)(1-\eta)\right] - \left[(1+\gamma)(1-\eta) - 1\right]\\ =&(1-\eta)\left[(1+\gamma)\log(1+\gamma) - \gamma\right] + (1+\gamma)(1-\eta)\log(1-\eta) + \eta\\ \geq&(1-\eta)\left[(1+\gamma)\log(1+\gamma) - \gamma\right] + (1+\gamma)(-\eta) + \eta\\ =&(1-\eta)\left[(1+\gamma)\log(1+\gamma) - \gamma\right] - \eta\gamma,\end{aligned}$$

where the last inequality is due to the fact that $(1-x)\log(1-x)+x \geq 0$ for any $0 < x < 1$ and $0 < \eta \leq \frac{\gamma}{4(1+\gamma)} < \frac{1}{4}$. Since $\log(1+\gamma) - \frac{\gamma}{1+\gamma} \geq \frac{\gamma^2}{2(1+\gamma)^2}$, we have $\eta\gamma \leq \frac{\gamma^2}{4(1+\gamma)} \leq \frac{1}{2}\left[(1+\gamma)\log(1+\gamma)-\gamma\right]$. Therefore,

$$\begin{aligned}\phi\left[(1-\eta)(1+\gamma)-1\right] \geq& (1-\eta)\left[(1+\gamma)\log(1+\gamma)-\gamma\right]-\eta\gamma \\ \geq& \left(\frac{1}{2}-\eta\right)\phi(\gamma) \geq \frac{1}{4}\phi(\gamma),\end{aligned}$$

where the last inequality is due to $0 < \eta \leq \frac{\gamma}{4(1+\gamma)} < \frac{1}{4}$. ■

C.4. Proof of Lemma 15

Lemma 15 *For binary entropy function $h(x) = -x\log x - (1-x)\log(1-x)$, $\phi(x) = (1+x)\log(1+x) - x$ and any constant $\delta \in (0,1)$, when $m \geq \log n$,*

$$\sum_{k=\delta m}^{m-1} \exp\left[-mh\left(\frac{k}{m}\right)\right] \leq (\log n)^{-1+o(1)}$$

Proof We note that

$$\begin{aligned}\sum_{k=\delta m}^{m-1} \exp\left[-mh\left(\frac{k}{m}\right)\right] &\leq \sum_{k=1}^{m-1} \exp\left[-mh\left(\frac{k}{m}\right)\right] \\ &\overset{(a)}{\leq} 2\sum_{1\leq k\leq \frac{m}{2}} \exp\left[-k\log\left(\frac{m}{k}\right)\right] \\ &\leq 2\sum_{1\leq k\leq 2\log m} \exp\left[-k\log\left(\frac{m}{k}\right)\right] + 2\sum_{2\log m<k\leq\frac{m}{2}} 2^{-k} \\ &\leq 2\cdot\exp\left(-\log m\right)\cdot\left(2\log m\right) + 2\cdot 2^{-2\log m} \overset{(b)}{\leq} (\log n)^{-1+o(1)},\end{aligned}$$

where (a) is because $h(x) = h(1-x)$ and $h(x) \geq -x\log x$ and (b) is because $m \geq \log n$. ■

C.5. Proof of Lemma 16

Lemma 16 *For $P \sim \mathrm{Bern}(p,p,\rho)$ and $Q \sim \mathrm{Bern}(p,p,0)$, the KL-divergence between P and Q can be upper bounded by:*

$$D(P\|Q) \leq 25p^2\phi(\gamma).$$

Proof By direct calculation,

$$\begin{aligned}
D(P\|Q) &= \sum_{\{a,b\}\in\{0,1\}} p_{ab}\log\left[\frac{p_{ab}}{p^{a+b}(1-p)^{2-a-b}}\right] \\
&= \left[p^2+\rho p(1-p)\right]\log\left[1+\frac{\rho(1-p)}{p}\right]+2p(1-p)(1-\rho)\log(1-\rho) \\
&+\left[(1-p)^2+\rho p(1-p)\right]\log\left(1+\frac{\rho p}{1-p}\right) \\
&\le \left[p^2+\rho p(1-p)\right]\log\left[1+\frac{\rho(1-p)}{p}\right]+2p(1-p)(1-\rho)\cdot(-\rho) \\
&+\left[(1-p)^2+\rho p(1-p)\right]\cdot\frac{\rho p}{1-p} \\
&= \left[p^2+\rho p(1-p)\right]\log\left[1+\frac{\rho(1-p)}{p}\right]-\rho p(1-p)+\rho^2\left[2p(1-p)+p^2\right].
\end{aligned}$$

Since $\log(1+x)\ge\frac{x}{x+1}+\frac{x^2}{2(x+1)^2}$ for any $x\ge 0$, we get $p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right]\ge\frac{p^2\gamma^2}{2(\gamma+1)}$. When $\gamma<3$, since $\frac{p^2\gamma^2}{2(\gamma+1)}\ge\frac{p^2\gamma^2}{8}=\frac{\rho^2(1-p)^2}{8}\ge\frac{\rho^2\left[2p(1-p)+p^2\right]}{24}$ for $0<p\le\frac{1}{2}$, we get

$$D(P\|Q)\le p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right]+\rho^2\left[2p(1-p)+p^2\right]\le 25p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right].$$

When $\gamma\ge 3$, since $p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right]\ge p^2\gamma(\log 4-1)=(\log 4-1)\rho p(1-p)$ and $\rho^2\left[2p(1-p)+p^2\right]\le 3\rho p(1-p)$, we get

$$\begin{aligned}
D(P\|Q) &\le p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right]+\rho^2\left[2p(1-p)+p^2\right] \\
&\le \left(\frac{3}{\log 4-1}+1\right)p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right]\le 25p^2\left[(1+\gamma)\log(1+\gamma)-\gamma\right].
\end{aligned}$$

Therefore, we get $D(P\|Q)\le 25p^2\phi(\gamma)$. ■

C.6. Proof of Lemma 17

Lemma 17 (Chernoff's inequality for Binomials) *Suppose* $\xi\sim\mathrm{Bin}(n,p)$, *denote* $\mu=np$, *then*

$$\mathbb{P}\left[\xi\ge(1+\delta)\mu\right]\le\exp\left\{-\mu\left[(1+\delta)\log(1+\delta)-\delta\right]\right\}, \tag{20}$$

$$\mathbb{P}\left[\xi\le(1-\delta)\mu\right]\le\exp\left(-\frac{\delta^2\mu}{2}\right). \tag{21}$$

We also have

$$\mathbb{P}\left[\xi\ge(1+\delta)\mu\right]\le\exp\left(-\frac{\delta\mu}{2+\delta}\right). \tag{22}$$

Proof By Theorems 4.4 and 4.5 in Mitzenmacher and Upfal (2005) we have (20) and (21). Since $(1+\delta)\log(1+\delta)-\delta\ge\frac{\delta^2}{2+\delta}$, we obtain (22) from (20). ■