On the Information-Theoretic Limits of Attributed Graph Alignment

by

Ning Zhang

B.Sc., Nankai University, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On the Information-Theoretic Limits of Attributed Graph Alignment

submitted by Ning Zhang in partial fulfillment of the requirements

for the degree of Master of Applied Science

in Electrical and Computer Engineering

Examining Committee:

Dr. Lele Wang, Assistant Professor, Electrical and Computer Engineering, UBC
Supervisor

Dr. Lutz Lampe, Professor, Electrical and Computer Engineering, UBC
Supervisory Committee Member

Dr. Cyril Leung, Professor, Electrical and Computer Engineering, UBC
Supervisory Committee Member

Abstract

Graph alignment aims at recovering the vertex correspondence between two correlated graphs, which is a task that frequently occurs in graph mining applications such as social network analysis, computational biology, etc. Existing studies on graph alignment mostly identify the vertex correspondence by exploiting the graph structure similarity. However, in many real-world applications, additional information attached to individual vertices, such as the user profiles in social networks, might be publicly available. In this thesis, we consider the attributed graph alignment problem, where additional information attached to individuals, referred to as attributes, is incorporated to assist graph alignment. We establish both the achievability and converse results on recovering vertex correspondence exactly, where the conditions match for a wide range of practical regimes. Our results span the full spectrum between models that only consider graph structure similarity and models where only attribute information is available.

Lay summary

Graphs, as the mathematical abstraction of entities and their relationships, widely appear in data science studies, such as social network analysis, computational biology, etc. In many of these applications, one may observe graph data about the same group of entities from multiple sources. For example, in social networks, individuals often maintain accounts on different platforms. The social networks obtained from different platforms often share some structural similarities, because they reflect the same underlying friendship network. Given a pair of such correlated graphs, a natural question to ask is: can we find the vertex correspondence between them?

In this thesis, we study the graph alignment method, a technique for finding the vertex correspondence between two correlated graphs. We propose a random graph model that generates correlated graph pairs with side information attributed to each vertex. Under our model, we characterize the information-theoretic limits of exactly aligning those graph pairs.

Preface

This thesis is a result of the joint work with Dr. Weina Wang and Dr. Lele Wang. Most of the contents come from the preprint [29]. A shorter version of this work is published in the *2021 IEEE International Symposium on Information Theory*.

Dr. Weina Wang and Dr. Lele Wang initially proposed the attributed graph alignment problem and formulated the attributed Erdős–Rényi pair model. They provided precious guidance on research directions as well as academic writing. My contribution to this work includes (1) establishing the parameter regime where exact attributed graph alignment can be achieved by the MAP estimator; (2) establishing the parameter regime where no algorithm guarantees exact alignment; and (3) specializing these regions in three closely related models: the Erdős–Rényi model, the bipartite model, and the seeded graph model.

Table of Contents

List of Figures

Notation

$\mathrm{Bin}(n, p)$	Binomial distribution with parameter $n \in \mathbb{N}$ and $p \in [0, 1]$
$\mathcal{E}_{\mathrm{u}}$	Set of user-user vertex pairs
$\mathcal{E}_{\mathrm{a}}$	Set of user-attribute vertex pairs
$f(n) = \omega(g(n))$	$\lim_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} = \infty$
$f(n) = o(g(n))$	$\lim_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} = 0$
$f(n) = O(g(n))$	$\limsup_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} < \infty$
$f(n) = \Omega(g(n))$	$\liminf_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} > 0$
$f(n) = \Theta(g(n))$	$f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$
$\mathrm{Hyp}(n, N, K)$	Hypergeometric distribution with parameter $n, N, K \in \mathbb{N}$
$\log(\cdot)$	Logarithm to the base of e
$\mathbb{N}$	Set of natural numbers
$[n]$	{1,2,. . . ,n}
$n!$	Factorial of the integer n
$\mathbb{R}$	Set of real numbers
R	Number of edges in the intersection graph
$\mathcal{S}_n$	Set of all permutations on n distinct elements
$\mathcal{S}_{n,\tilde{n}}$	Set of all permutations on n distinct elements with $\tilde{n}$ fixed points
$\mathcal{V}_{\mathrm{u}}$	User vertex set of a graph
$\mathcal{V}_{\mathrm{a}}$	Attribute vertex set of a graph
$\Delta^{\mathrm{u}}(g_1, g_2)$	The number of user-user edge disagreements between graph g_1 and g_2
$\Delta^{\mathrm{a}}(g_1, g_2)$	The number of user-attribute edge disagreements between graph g_1 and g_2
Π^*	The underlying true permutation on the set $[n]$
π_{id}	The identity permutation
ρ	Correlation coefficient

Acknowledgements

First and foremost, I would like to extend my gratitude to my supervisor Dr. Lele Wang, who introduced me to this exciting area about graphs and high dimensional statistics. She provided me with much support and is always available for answering my questions. Her enthusiasm for research has inspired me during the entire Master's program. This thesis would not have been possible without her help.

I owe particular thanks to Dr. Weina Wang for all the critical help and instructions on this project. Her insights helped me through many bottlenecks, and her intriguing feedback always pushed me to think a few steps further. Working with her not only improved my problem-solving skills, but more importantly, shaped my way of conducting research.

I am grateful to Dr. Lutz Lampe and Dr. Cyril Leung for their time serving on my thesis committee and their constructive feedback.

I feel lucky to learn from many brilliant researchers at UBC. I took Dr. Joel Friedman's course about spectral methods and was particularly impressed by how he connected ideas among several different research areas. I enjoyed learning optimization theories in the course instructed by Dr. Christos Thrampoulidis. Thanks to Dr. Renjie Liao's informative feedback and remarks during the reading group, I gained a deeper understanding of probabilistic machine learning. I thank Dr. Shuo Tang for her understanding and encouragement, which gave me the courage to make a huge transition in my research direction.

I thank MSD group members Jiaming Chen, Qi Yan, Animesh Sakorikar and Ziao Wang for many discussions and help. I would also like to extend my gratitude to all of my friends for their genuine help and encouragement. I feel lucky to have their accompany, either in person or online.

I am also grateful to Dr. Xin Chen at the University of Nottingham for his continuous support and many important help. I thank Dr. Fang Bo and Dr. Chuanyong Li at Nankai University for giving me many suggestions on research methodology and my future career.

Finally, special thanks are owed to my parents and sister for their unconditional love and support.

Chapter 1

Introduction

1.1 Motivation

The graph alignment problem, also known as the graph matching problem or the noisy graph isomorphism problem, has received growing attention in recent years, brought into prominence by applications in a wide range of areas [1, 13, 25]. For instance, in social network de-anonymization [15, 21], one is given two graphs, each of which represents the user relationship in a social network (e.g., Twitter, Facebook, Flickr, etc). One graph is anonymized and the other graph has user identities as public information. Then the graph alignment problem, whose goal is to find the best correspondence of two graphs with respect to a certain criterion, can be used to de-anonymize users in the anonymous graph by finding the correspondence between them and the users with public identities in the other graph.

The graph alignment problem has been studied under various random graph models, among which the most popular one is the *Erdős–Rényi graph pair* model (see, e.g., [4, 23, 27]). In particular, two Erdős–Rényi graphs on the same vertex set, G_1 and G_2, are generated in a way such that their edges are correlated. Then G_1 and an anonymous version of G_2, denoted as G_2', are made public, where G_2' is modeled as a vertex-permuted G_2 with an unknown permutation. Under this model, typically the goal is to achieve the so-called *exact alignment*, i.e., recovering the unknown permutation and thus revealing the correspondence for all vertices exactly.

A fundamental question in the graph alignment problem is: *when is exact alignment possible?* More specifically, *what conditions on the statistical properties of the graphs are required for achieving exact alignment when given unbounded computational resources?* Such conditions, usually referred to as *information-theoretic limits*, have been established for the Erdős–Rényi graph pair in a line of work [4, 5, 23, 27]. The best known information-theoretic limits are proved in [5, 27], where the authors establish nearly matching achievability and converse bounds.

In many real-world applications, additional information about the anonymized vertices might be available. For example, Facebook has user profiles on

their website about each user's age, birthplace, hobbies, etc. Such associated information is referred to as attributes (or features), which, unlike user identities, are often publicly available. Then a natural question to ask is: *Can the attribute information help recover the vertex correspondence?* If so, *can we quantify the amount of benefit brought by the attribute information?* The value of attribute information has been demonstrated in the work of aligning Netflix and IMDb users by [22]. They successfully recovered some of the user identities in the anonymized Netflix dataset based only on users' ratings of movies, without any information on the relationship among users. In this thesis, we incorporate attribute information to generalize the graph alignment problem. We call this problem the *attributed graph alignment* problem.

To investigate the attributed graph alignment problem, we extend the current Erdős–Rényi graph pair model and we refer to this new random graph model as the attributed Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. For a pair of graphs, G_1 and G_2, generated from the attributed Erdős–Rényi pair model, each graph contains n *user vertices* and m *attribute vertices* (see Figure 1.1). Here, the user vertices represent the entities that need to be aligned; while the attribute vertices are all pre-aligned, reflecting the public availability of the attribute information. There are two types of edges in each graph, i.e., edges between user vertices and edges between user vertices and attribute vertices. Here, edges between user vertices represent the relationship between users (e.g., friendship relations in a social network); edges between user vertices and attribute vertices encode the side information attached to each user (e.g., user profiles in a social network). These two types of edges are correlatedly generated in the following way: for a user-user vertex pair (i, j), the edges connecting them follow a distribution $\boldsymbol{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, where p_{11} is the probability that i and j are connected in both G_1 and G_2, and p_{10}, p_{01}, p_{00} represent the three remaining cases respectively: i, j are only connected in G_1, only connected in G_2, and not connected in neither G_1 nor G_2; for a user-attribute vertex pair, the edges connecting them are generated in a similar way following a distribution $\boldsymbol{q} = (q_{11}, q_{10}, q_{01}, q_{00})$. This random process creates an identically labeled graph pair (G_1, G_2) with similarity in both the graph topology part (user-user edges) and the attribute part (user-attribute edges). The graph G_2 is then anonymized by applying a random permutation on its *user vertices* and the anonymized graph is denoted as G_2'. Under this formulation, our goal of attributed graph alignment is to recover this unknown permutation from G_1 and G_2' by exploring both the topology similarity and attribute similarity.

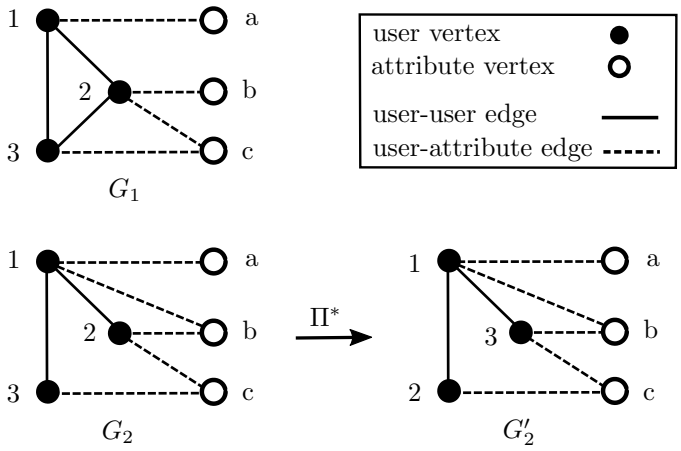

Figure 1.1: **An example of the attributed Erdős–Rényi graph pair.** Graph G_1 and G_2 are generated on the same set of vertices. The anonymized graph G_2' is obtained by applying $\Pi^* = (1)(2,3)$ only on $\mathcal{V}_\mathrm{a}$ of G_2 (permutation Π^* is written in cycle notation).

1.2 Results Overview

Under our attributed Erdős–Rényi pair model, we use the maximum a posterior (MAP) estimator for aligning (G_1, G_2'), and establish the achievability and converse results for exact alignment. To get an intuitive understanding of how the existence of attribute information contributes to exact graph alignment, we present a simplified result by considering a special case that is typical and interesting in practice. We defer the general result to Section 2.1. In most social networks, the degree of a vertex is much smaller than the total number of users. Based on this observation, we assume that the marginal edge probabilities are bound away from 1, i.e., $1 - (p_{11} + p_{10}) = \Theta(1)$ and $1 - (p_{11} + p_{01}) = \Theta(1)$. In addition, two social networks on the same set of users are normally highly correlated. Based on this, we assume that the correlation coefficient of user-user edges, denoted as ρ_u, is not vanishing, i.e., $\rho_\mathrm{u} = \Theta(1)$. Moreover, we assume that the number of attributes satisfies $m = \Omega((\log n)^3)$. Under these three assumptions, we establish the following asymptotically matching achievability and converse result as $n \to \infty$.

- **Achievability:** If

$$np_{11} + mq_{11} - \log n \to \infty, \tag{1.1}$$

  then there exists an algorithm that achieves exact alignment *with high probability* (w.h.p.)

- **Converse:** If

$$np_{11} + mq_{11} - \log n \to -\infty, \tag{1.2}$$

  then no algorithm guarantees exact alignment w.h.p.

Here np_{11} is the average number of common users between G_1 and G_2 that are connected to a identical user vertex, and mq_{11} is the average number of common attributes. Intuitively, the key quantity $np_{11} + mq_{11}$ (average common vertex degree) quantifies the topology and attribute similarity between G_1 and G_2. The above results simply show that if this similarity measure is large enough, then exact alignment is achievable, or otherwise no algorithm can exactly recover the true alignment. It is also worth noticing that the average common vertex degree in attribute, i.e., mq_{11} out of the overall average common vertex degree highlights the extra benefit from attribute information. The achievability and converse results are illustrated in Figure 1.2.

From the information-theoretic limits we derive for the attributed graph pair, we could obtain information-theoretic limits on other existing random graph models as special cases (see Figure 1.2). Here, we highlight how our results, by comparing with the three specialized settings, help answer some of the existing problems in the graph alignment literature. We defer a more detailed comparison to Chapter 2.2.

- Specializing our model by setting $q_{00} = 1$, we remove the effect of the attribute vertices and get the correlated Erdős–Rényi graph pair model. This specialized result recovers the information-theoretic limits on Erdős–Rényi graph alignment in [5, 27]. Comparing the specialized and unspecialized results allows us to quantify the benefit brought by the attribute information.

- Specializing our model by setting $\boldsymbol{p} = \boldsymbol{q}$, we can then treat the m attribute vertices as pre-aligned user vertices and get the seeded Erdős–Rényi model. Our specialized results provide information-theoretic limits on the seeded graph alignment problem.

- Specializing our model by setting $p_{00} = 1$, we remove all of the user-user edges and get the correlated bipartite graph pair model. The specialized results provide information-theoretic limits on bipartite graph alignment [3, 24].

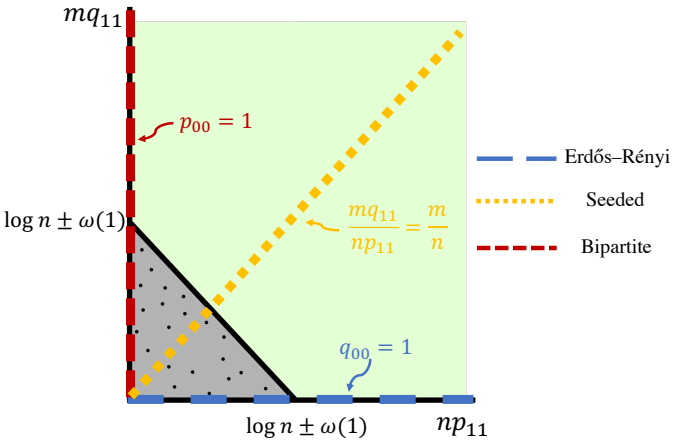

Figure 1.2: **Simplified information-theoretic limits on the attributed Erdős–Rényi graph pair alignment.** The green region in the figure is information-theoretically achievable, which is specified by condition (1.1); the shaded grey region is not achievable by any algorithm and is specified by condition (1.2). The three lines represent three specialized settings respectively: (1) the blue line, which corresponds to the correlated Erdős–Rényi model, is obtained by setting $q_{00} = 1$; (2) the yellow line, which corresponds to the seeded Erdős–Rényi model, is obtained by setting $\boldsymbol{p} = \boldsymbol{q}$ (correspondingly $p_{11} = q_{11}$ in the figure); (3) the red line, which corresponds to the correlated bipartite model, is obtained by setting $p_{00} = 1$ The intersections of the lines with the achievable and not achievable regions give the information-theoretic limits of the correlated Erdős–Rényi model, seeded Erdős–Rényi model and the correlated bipartite model respectively.

1.3 Model Formulation

In this section, we describe the attributed Erdős–Rényi graph pair model. Under this model formulation, we formally define the exact attributed graph alignment problem. An illustration of the model is given in Figure 1.1.

User vertices and attribute vertices. We first generate two graphs, G_1 and G_2, on the same vertex set $\mathcal{V}$. The vertex set $\mathcal{V}$ consists of two disjoint sets of vertices, the *user vertex set* $\mathcal{V}_{\mathrm{u}}$ and the *attribute vertex set* $\mathcal{V}_{\mathrm{a}}$, i.e., $\mathcal{V} = \mathcal{V}_{\mathrm{u}} \cup \mathcal{V}_{\mathrm{a}}$. Assume that the user vertex set $\mathcal{V}_{\mathrm{u}}$ consists of n vertices, labeled as $[n] \triangleq \{1, 2, 3, ..., n\}$. Assume that the attribute vertex set $\mathcal{V}_{\mathrm{a}}$ consists of m vertices, and m scales as a function of n.

Correlated edges. To describe the probabilistic model for edges in G_1 and G_2, we first consider the set of user-user vertex pairs $\mathcal{E}_{\mathrm{u}} \triangleq \mathcal{V}_{\mathrm{u}} \times \mathcal{V}_{\mathrm{u}}$ and the set of user-attribute vertex pairs $\mathcal{E}_{\mathrm{a}} \triangleq \mathcal{V}_{\mathrm{u}} \times \mathcal{V}_{\mathrm{a}}$. Then for each vertex pair $e \in \mathcal{E} \triangleq \mathcal{E}_{\mathrm{u}} \cup \mathcal{E}_{\mathrm{a}}$, we write $G_1(e) = 1$ (resp. $G_2(e) = 1$) if there is an edge connecting the two vertices in the pair in G_1 (resp. G_2), and write $G_1(e) = 0$ (resp. $G_2(e) = 0$) otherwise. Since we often consider the same vertex pair in both G_1 and G_2, we write $(G_1, G_2)(e)$ as a shortened form of $(G_1(e), G_2(e))$.

The edges of G_1 and G_2 are then correlatedly generated in the following way. For each user-user vertex pair $e \in \mathcal{E}_{\mathrm{u}}$, $(G_1, G_2)(e)$ follows the joint distribution specified by

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } p_{11}, \\ (1,0) & \text{w.p. } p_{10}, \\ (0,1) & \text{w.p. } p_{01}, \\ (0,0) & \text{w.p. } p_{00}, \end{cases} \tag{1.3}$$

where $p_{11}, p_{10}, p_{01}, p_{00}$ are probabilities that sum up to 1. For each user-attribute vertex pair $e \in \mathcal{E}_{\mathrm{a}}$, $(G_1, G_2)(e)$ follows the joint probability distribution specified by

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } q_{11}, \\ (1,0) & \text{w.p. } q_{10}, \\ (0,1) & \text{w.p. } q_{01}, \\ (0,0) & \text{w.p. } q_{00}, \end{cases} \tag{1.4}$$

where $q_{11}, q_{10}, q_{01}, q_{00}$ are the probabilities and they sum up to 1. The correlation between $G_1(e)$ and $G_2(e)$ is measured by the correlation coefficient defined as

$$\rho(e) \triangleq \frac{\mathsf{Cov}(G_1(e), G_2(e))}{\sqrt{\mathsf{Var}[G_1(e)]}\sqrt{\mathsf{Var}[G_2(e)]}},$$

where $\mathsf{Cov}(G_1(e), G_2(e))$ is the covariance between $G_1(e)$ and $G_2(e)$ and $\mathsf{Var}[G_1(e)]$ and $\mathsf{Var}[G_2(e)]$ are the variances. We assume that $G_1(e)$ and $G_2(e)$ are positively correlated, i,e., $\rho(e) > 0$ for every vertex pair e. Across different vertex pair e's, the $(G_1, G_2)(e)$'s are independent. Finally, remember that there are no edges between attribute vertices in our model.

For compactness of notation, we represent the joint distributions in (2.12) and (2.30) in the following matrix form:

$$\boldsymbol{p} = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} \quad \text{and} \quad \boldsymbol{q} = \begin{pmatrix} q_{11} & q_{10} \\ q_{01} & q_{00} \end{pmatrix}.$$

We refer to the graph pair (G_1, G_2) as an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}, m, \boldsymbol{q})$. Note that this model is equivalent to the subsampling model in the literature [23].

Anonymization and exact alignment. In the attributed graph alignment problem, we are given G_1 and an anonymized version of G_2, denoted as G_2'. The anonymized graph G_2' is generated by applying a random permutation Π^* on the user vertex set of G_2, where the permutation Π^* is unknown. More explicitly, each user vertex i in G_2 is re-labeled as $\Pi^*(i)$ in G_2'. The permutation Π^* is chosen uniformly at random from $\mathcal{S}_n$, where $\mathcal{S}_n$ is the set of all permutations on $[n]$. Since G_1 and G_2' are observable, we refer to (G_1, G_2') as the *observable pair* generated from the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}, m, \boldsymbol{q})$.

Then the graph alignment problem, i.e., the problem of recovering the identities/original labels of user vertices in the anonymized graph G_2', can be formulated as a problem of estimating the underlying permutation Π^*. The goal of graph alignment is to design an estimator $\hat{\pi}(G_1, G_2')$ as a function of G_1 and G_2' to best estimate Π^*. We say $\hat{\pi}(G_1, G_2')$ achieves *exact alignment* if $\hat{\pi}(G_1, G_2') = \Pi^*$. The probability of error for exact alignment is defined as $\mathsf{P}(\hat{\pi}(G_1, G_2') \neq \Pi^*)$. We say that exact alignment is achievable w.h.p. if there exists $\hat{\pi}$ such that $\lim_{n\to\infty} \mathsf{P}(\hat{\pi}(G_1, G_2') \neq \Pi^*) = 0$.

1.4 Related Work

The exact graph alignment problem has been studied under various random graph models. One of the most popular models is the correlated Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p})$, which generates simple graph pairs without any side information. Under this model, the optimal alignment strategy, derived from the MAP estimator, is enumerating all possible permutations in order to make the two graph achieve the maximum edge overlap. While the

optimal strategy is NP-hard, numerous studies have proposed polynomial-time approaches that exactly solve the graph alignment problem with high probability [7–9, 19]. Here, we do not attempt to provide further detailed discussions on efficient algorithms, but focus on surveying the information-theoretic limits of exact alignment. Currently, the best-known information-theoretic limits on Erdős–Rényi graph alignment are shown in [5] and [27] by analyzing error event of the MAP estimator. In particular, both [5] and [27] prove achievability in the regime $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq (2+\epsilon)\log n$ using a combinatorial method called cycle decomposition. In [5], the authors further prove that under some sparsity assumptions, the achievability result can be improved to $np_{11} \geq \log n + \omega(1)$; in [27], the authors prove that in the symmetric case where $p_{10} = p_{01}$, the achievability result can be extended to $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq (1+\epsilon)\log n$. Conversely, [5] proves that exact alignment is not achievable if $np_{11} \leq \log n - \omega(1)$ by showing the existence of isolated vertices in the intersection graph $G_1 \cap G_2$; [27] expands the converse region to $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \leq (1-\epsilon)\log n$ by showing the existence of permutations that can fail the MAP estimator by swapping two vertices. From the aforementioned results, we can see that there is a gap between the achievability and the converse. Closing this gap is still an open problem in the field.

Recently, there has been a growing interest in studying graph alignment with side information. For example, in the seeded alignment setting, the side information appears in the form of a partial observation of the latent alignment. For the seeded graph alignment problem, there have been a number of studies concentrating on designing polynomial-time algorithms with performance guarantees [10, 18, 20]. Some other more general settings treat any form of side information as vertex attributes and formulate this as the attribute graph alignment problem [31]. There is a line of empirical studies on the attributed graph alignment [30–32], yet, to the best of our knowledge, there is no known result on information-theoretic limits on graph alignment with attribute information.

1.5 Summary of Contributions

The main contributions of this thesis can be summarized as follows.

1. ***Model Formulation.*** We propose the attributed Erdős–Rényi pair model, which incorporates both the graph topology similarly and the attribute similarity. Such model formulation allows us to align graphs with the assistance of publicly available side information. Moreover, our model

serves as a unifying formulation in the graph alignment literature and includes the correlated Erdős–Rényi model, seeded Erdős–Rényi model and the correlated bipartite model as its special cases.

2. ***Information theoretic limits.*** We establish the achievability and converse results on exactly aligning random attributed graphs, which coincide under assumptions that are typical and interesting in practice. Our results span the full spectrum from the traditional Erdős–Rényi pair model where only the user relationship networks are available to models where only attribute information is available, unifying the existing results in each of these settings.

3. ***Proof techniques.*** The proof techniques for the achievability results are mainly inspired by the previous study on Erdős–Rényi graph alignment [5]. For the converse result, our proof of the converse first studies the sharp phase-transition phenomenon on the existence of *indistinguishable vertex pairs*, which may of independent interest in research about random graphs [11].

1.6 Thesis Outline

In this Introduction chapter, we give an overview of our problem formulation, present a simplified version of our results, and summarize our contributions. The rest of this thesis is organized as follows: our main results are presented in Chapter 2. In Chapter 2.1 we present both the achievability and converse results on the exact attributed graph alignment. Then, in Chapter 2.2, we specialize our results into three closely related random graph models, and compare our specialized results with the best-known results in the literature. In Chapter 3, we present the detailed proof of the achievability results. In particular, we derive the general achievability result in Chapter 3.1 and prove the achievability of the sparse region in Chapter 3.2. In Chapter 4 , we prove the converse result. In Chapter 5, we summarize several extensions and potential future directions of our work.

Chapter 2

Main Result

In Chapter 2.1, we present the achievability and the converse results. Our achievability results characterize the parameter regime of the attributed Erdős–Rényi model, under which there exists an algorithm that guarantees exact alignment with high probability; while the converse result shows the parameter regime, under which no algorithm achieves exact alignment with high probability. Here, our analysis of both the achievability and the converse is based on the assumption that we have access to unbounded computational power and do not require the algorithms to be computationally efficient. In Chapter 2.2, we compare our specialized results with the best-known results on three well-studied models in the graph alignment literature.

2.1 Achievability and Converse

In this section, we state the general achievability results (Theorem 1 and Theorem 2) and the converse results (Theorem 3). To better demonstrate the benefit from attribute information, we also present a simplified version of the result under some mild and practical assumptions as Corollary 1.

For compactness when present our results, here we define $\psi_{\mathrm{u}} \triangleq (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2$ and $\psi_{\mathrm{a}} \triangleq (\sqrt{q_{11}q_{00}} - \sqrt{q_{10}q_{01}})^2$. We will use these notations throughout the thesis.

Theorem 1 (General achievability)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If*

$$\frac{n\psi_{\mathrm{u}}}{2} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{2.1}$$

then the MAP estimator achieves exact alignment w.h.p.

Theorem 2 (Achievability in sparse region)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If*

$$p_{11} = O\left(\frac{\log n}{n}\right), \tag{2.2}$$

$$p_{10} + p_{01} = O\left(\frac{1}{\log n}\right), \tag{2.3}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\frac{1}{(\log n)^3}\right), \tag{2.4}$$

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{2.5}$$

then the MAP estimator achieves exact alignment w.h.p.

Theorem 3 (Converse)**.** *Consider the attributed Erdős–Rényi pair* $\mathcal{G}(n, \boldsymbol{p}, m, \boldsymbol{q})$*. If*

$$np_{11} + mq_{11} - \log n \to -\infty, \tag{2.6}$$

then no algorithm guarantees exact alignment w.h.p.

To better illustrate the benefit of attribute information in the graph alignment problem, we present in Corollary 1 a simplified version of our achievability result by adding mild assumptions on user-user edges motivated by practical applications. This simplified result also makes it easier to compare the achievability result to the converse result in Theorem 3, which will be illustrated in Figure 1.2.Note that these additional assumptions are not needed for technical proofs.

In a typical social network, the degree of a vertex is much smaller than the total number of users. Based on this observation, we assume that the marginal probabilities of an edge in both G_1 and G_2 are not going to 1, i.e.,

$$1 - (p_{11} + p_{10}) = \Theta(1), \quad 1 - (p_{11} + p_{01}) = \Theta(1). \tag{2.7}$$

In addition, two social networks on the same set of users are typically highly correlated. Based on this, we assume that the correlation coefficient of user-user and user-attribute edges, denoted as ρ_{u} and ρ_{a} , are not vanishing, i.e.,

$$\rho_{\mathrm{u}} = \Theta(1), \quad \rho_{\mathrm{a}} = \Theta(1). \tag{2.8}$$

Corollary 1 (Simplified achievability)**.** Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ under conditions (2.7) and (2.8). If

$$np_{11} + m\psi_{\mathrm{a}} - \log n \to \infty \tag{2.9}$$

then there exists an algorithm that achieves exact alignment *w.h.p.*
If we further have $m = \Omega((\log n)^3)$, then the above condition (2.9) becomes

$$np_{11} + mq_{11} - \log n \to \infty, \tag{2.10}$$

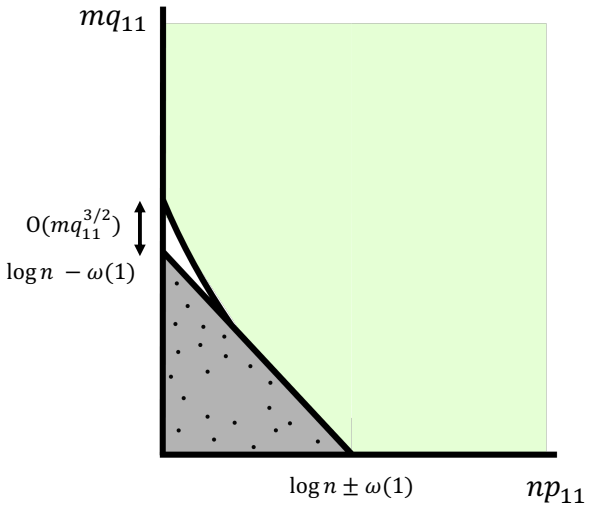

Figure 2.1: **Simplified information-theoretic limits on the attributed Erdős–Rényi graph pair alignment.** The green region is information theoretically achievable and is specified by condition (2.11); the shaded grey region in not achievable according the Theorem 3. The gap between the achievability and converse represent $q_{11} - \psi_{\mathrm{a}} = O(q_{11}^{3/2})$. In particular, this gap is negligible up to $\pm\omega(1)$ when $m = \Omega((\log n)^3)$, which is the other simplified case presented in Figure 1.2.

If $m = o((\log n)^3)$, the above condition (2.9) becomes the following.

$$np_{11} + mq_{11} - ma_n - \log n \to \infty, \tag{2.11}$$

where $a_n = q_{11} - \psi_{\mathrm{a}} = O(q_{11}^{3/2})$.

This simplified achievability and the converse results are visualized in Figure 2.1. The gap between the achievability and the converse comes from the difference between mq_{11} and $m\psi_{\mathrm{a}}$. This gap is negligible up to $\pm\omega(1)$ when $m = \Omega((\log n)^3)$ (Figure 1.2). Closing the gap for the $m = o((\log n)^3)$ case remains an open problem.

2.2 Specialization and Comparison

In this section, we specialize our main results (Theorem 1, Theorem 2 and Theorem 3) on exact alignment of attributed Erdős–Rényi pair model and compare them with the best-known results from three closely related graph

alignment topics: the Erdős–Rényi graph alignment, the seeded Erdős–Rényi graph alignment and the bipartite graph alignment.

2.2.1 Erdős–Rényi graph pair

The correlated Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p})$ is the most commonly studied setting for graph alignment tasks that consider only graph topology similarity [4, 5, 7, 23, 27]. This model generates graph pairs that contain user vertices only. For a pair of graphs G_1, G_2 obtained from this model $\mathcal{G}(n, \boldsymbol{p})$, we use $\mathcal{V}_{\mathrm{u}}$ to denote their vertex set and $|\mathcal{V}_{\mathrm{u}}| = n$. The edges in G_1 and G_2 are jointly generated in the following way: for a pair of users $e \in \binom{\mathcal{V}_{\mathrm{u}}}{2}$, we have

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } p_{11}, \\ (1,0) & \text{w.p. } p_{10}, \\ (0,1) & \text{w.p. } p_{01}, \\ (0,0) & \text{w.p. } p_{00}. \end{cases} \tag{2.12}$$

The anonymized graph G_2' is obtained by applying a random permutation Π^* on the vertices of G_2 directly. This model can be specialized from the attributed graph pair model by setting the number of attributes $m = 0$. For aligning such correlated Erdős–Rényi pair, the best-known information-theoretic limits are established in [5, 27] and we state the combined results here for ease of comparison.

Theorem 4 ([5, 27])**.** *Consider the correlated Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{q})$.*
Achievability:
If

$$n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq 2\log n + \omega(1), \tag{2.13}$$

or

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{2.14}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{2.15}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{2.16}$$

$$np_{11} \geq \log n + \omega(1), \tag{2.17}$$

then the MAP estimator achieves exact alignment w.h.p.
Converse:
If there exists a constant ϵ such that

$$n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \leq (1-\epsilon)\log n,$$

or

$$np_{11} \leq \log n - \omega(1),$$

then no algorithm guarantees exact alignment w.h.p.

To compare our results with the best-known information-theoretic limits on Erdős–Rényi graph alignment, we first specialize our model to fit the setting in the correlated Erdős–Rényi pair model. The attributed Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ degenerate to the Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p})$ by removing the attributed vertices. As a consequence of the specialization, we directly obtain the following achievability and converse result from Theorem 1, Theorem 2 and Theorem 3. Comparing the specialized results (Theorem 5) and the best-known results (Theorem 4), we can see that the specialized achievability is the same as the best-known achievability, while the specialized converse is strictly contained by the best-known results in Theorem 4.

Theorem 5 (Specialization to the correlated Erdős–Rényi pair model)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; 0, \boldsymbol{q})$.*
Achievability: If

$$n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq 2\log n + \omega(1), \tag{2.18}$$

or

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{2.19}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{2.20}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{2.21}$$

$$np_{11} = \log n + \omega(1), \tag{2.22}$$

then the MAP estimator achieves exact alignment w.h.p.
Converse: If

$$np_{11} \leq \log n - \omega(1),$$

then no algorithm guarantees exact alignment w.h.p.

2.2.2 Seeded Erdős–Rényi pair

In the seeded graph model $\mathcal{G}(n, m, \boldsymbol{p})$, a pair of graphs G_1, G_2 are generated from the correlated Erdős–Rényi pair model $\mathcal{G}(n+m, \boldsymbol{p})$. Then the anonymized graph G_2' is obtained by applying a random permutation on the vertices of G_2. In addition to knowing G_1 and G_2', in the seeded graph setting, we are also given the true alignment on part of the user vertices, which is known as the seed set $\mathcal{V}_s$. The number of aligned pairs in $\mathcal{V}_s$ is a fixed number m. The seeded alignment problem has been studied by [15, 17, 18, 20, 28][1]. To the best of our knowledge, the best information-theoretic limits of the seeded alignment problem are given by [20, 26].

Theorem 6 (Best-known information-theoretic limits [20, 26])**.** *Consider the seeded Erdős–Rényi graph pair $\mathcal{G}(n, m, \boldsymbol{p})$.*
Achievability from [20]: Assume that $p_{10} = p_{01}$, $\frac{p_{11}}{p_{11}+p_{10}} = \Theta(1)$ and $(n+m)p_{11} - \log(n+m) \to \infty$.

1. *Suppose that for a fixed constant $\epsilon > 0$, we have*

$$(n+m)\frac{(p_{11}+p_{10})^2}{p_{11}} \le (n+m)^{1/2-\epsilon},$$
$$\frac{m}{m+n} \ge (n+m)^{-1/2+3\epsilon}.$$

   Then Algorithm 1 in [20] achieves exact alignment with probability at least $1 - o(1)$.

2. *Here, we define the constants $a, b \in (0, 1]$ as*

$$(n+m)\frac{(p_{11}+p_{10})^2}{p_{11}} = b(n+m)^a.$$

   Let $d = \left\lfloor \frac{1}{a} \right\rfloor + 1$ and $s = \frac{p_{11}}{p_{11}+p_{00}}$. Suppose that

$$b \le \frac{s}{16(2-s)^2} \quad \text{and} \quad \frac{m}{m+n} \ge \frac{300\log(n+m)}{((n+m)p_{11})^{d-1}}.$$

   Then Algorithm 2 in [20] achieves exact alignment with probability at least $1 - 4(n+m)^{-1}$.

[1] In the literature, both random [15] and deterministic [28] seed sets are considered. Here, we focus on the deterministic seed set setting which is closely related to our attributed Erdős–Rényi pair model.

3. *Suppose for a fixed constant* $\epsilon < 1/6$*, we have*

$$(n+m)\frac{(p_{11}+p_{10})^2}{p_{11}} \le (n+m)^{\epsilon},$$
$$\frac{m}{m+n} \ge (n+m)^{-1+3\epsilon}.$$

Then Algorithm 3 in [20] achieves exact alignment with probability at least $1 - o(1)$*.*

Achievability from [26] Assume that $p_{10} = p_{01}$*,* $\frac{p_{11}}{p_{11}+p_{10}} = \Theta(1)$*, and* $\frac{(p_{11}+p_{10})^2}{p_{11}} = o(1)$*.*

1. *In the regime where* $mp_{11} = \Omega(\log n)$*, if for a constant* $\epsilon > 0$*, we have*

$$(n+m)p_{11} \ge (1+\epsilon)\log n,$$

then the ATTRRICH *algorithm in [26] achieves exact alignment w.h.p.*

2. *In the regime where* $mp_{11} = o(\log n)$*, if for a constant* $\tau > 0$*, we have*

$$np_{11} - \log n = \omega(1),$$
$$mp_{11} \ge \frac{2\log n}{\tau \log(p_{11}/(p_{11}+p_{10})^2)},$$

then the ATTRSPARSE *algorithm in [26] achieves exact alignment w.h.p.*

Converse from [20] Consider the seeded Erdős–Rényi graph pair $\mathcal{G}(n, m, \boldsymbol{p})$*. If*

$$(n+m)p_{11} \le \log(n+m) + O(1), \quad \textit{and} \quad m = O(n),$$

then any algorithm fails with probability at least $\Theta(1)$*.*

To compare the best-known information-theoretic limits of the seeded Erdős–Rényi alignment with our results, we specialize the attributed Erdős–Rényi pair model by setting $\boldsymbol{p} = \boldsymbol{q}$, where m attribute vertices are pre-aligned seeds. Notice that a small difference between the $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{p})$ model and the seeded model $\mathcal{G}(n, m, \boldsymbol{p})$ is that there are no edges between the seeds in the specialized model but those edges exist in the seeded model. Such distinction may lead to a difference in the design of seeded graph alignment algorithms (e.g. algorithms from [20] used seed-seed edges). It turns out that such seed-seed edges have no influence on the optimal MAP estimators for the two models, which leads to the next lemma.

Lemma 1. *The information-theoretic limits on aligning the seeded Erdős–Rényi pair model $\mathcal{G}(n, m, \boldsymbol{p})$ and the information-theoretic limits on aligning the specialized attributed Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{p})$ are identical.*

Proof. See Lemma 6in the Appendix. □

Based on Lemma 1, we directly obtain the achievability and converse results on seeded graph alignment from Theorems 1, 2, and 3 by setting $\boldsymbol{p} = \boldsymbol{q}$. The specialized achievability and converse results strictly improve the best known achievability and converse in the literature for seeded graph alignment.

Theorem 7 (Specialization to the seeded Erdős–Rényi pair model)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{p})$.*
Achievability: If

$$(n+m)(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq 2\log n + \omega(1), \tag{2.23}$$

or

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{2.24}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{2.25}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{2.26}$$

$$np_{11} + m(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 = \log n + \omega(1), \tag{2.27}$$

then the MAP estimator achieves exact alignment w.h.p.
Converse: If

$$(n+m)p_{11} \leq \log n - \omega(1),$$

then no algorithm guarantees exact alignment w.h.p.

In particular, for seeded graph pairs that satisfy the two assumptions on typical social networks (condition 2.7 and (2.8) in Chapter 2.1), we have the following achievability and converse results.

Corollary 2. Consider the seeded Erdős–Rényi pair model $\mathcal{G}(n, m, \boldsymbol{p})$ under conditions (2.7) and (2.8).
Achievability: If

$$(n+m)p_{11} \geq \log n + \omega(1), \tag{2.28}$$

then the MAP estimator achieves exact alignment w.h.p.
Converse: If

$$(n+m)p_{11} \leq \log n - \omega(1), \tag{2.29}$$

then no algorithm guarantees exact alignment w.h.p.

Remark 1. In Corollary 2, we obtain asymptotically tight achievability and converse results under conditions (2.7) and (2.8).
Comparison between the achievability results: The achievability result in Corollary 2 strictly improve the best-known achievability results for seeded alignment [20, 26]. In the following, we provide a comparison between the achievability results in Corollary 2 and Theorem 6.

- For algorithms in [20], they all require

  $$(n+m)p_{11} \geq \log(n+m) + \omega(1),$$

  which is strictly contained by achievability condition (2.28)

  $$(n+m)p_{11} \geq \log n + \omega(1).$$

  To illustrate the strict inclusion, consider the case where

  $$p_{11} = \frac{\log(n+m)}{n+m}, \;\; p_{10} = p_{01} = 0, \;\text{ and }\; m = n^2.$$

  Then we have $(n+m)p_{11} = \log(n+m) < \log(n+m)+\omega(1)$, which is *not* feasible using algorithms from [20]. However, for such choices of p_{11} and m, we have $(n+m)p_{11} = \log(n+m) = \log(n+n^2) \geq \log n + \omega(1)$, which is feasible according to condition (2.28). Moreover, we have $1-(p_{11}+p_{01}) = 1-(p_{11}+p_{10}) = 1-\frac{\log(n+m)}{n+m} = \Theta(1)$ and $\rho_{\mathrm{u}} = \rho_{\mathrm{a}} = 1$, satisfying conditions (2.7) and (2.8). For later examples illustrating the strict inclusion, they all satisfy conditions (2.7) and (2.8), which can be checked following the same verification procedure. In those examples, we do not repeat such verification steps for compactness.

- For algorithms in [26], the ATTRRICH algorithm requires that

  $$(n+m)p_{11} \geq (1+\epsilon)\log n,$$

  which is a strict subset of condition (2.28). To illustrate the strict inclusion, consider the case where

  $$p_{11} = \frac{\log n + \log\log n}{n+m} \;\text{ and }\; p_{10} = p_{01} = 0.$$

Then we have $(n+m)p_{11} = \log n + \log\log n < (1+\epsilon)\log n$, which is *not* feasible using the ATTRRICH algorithm. However, for such choice of p_{11}, we have $(n+m)p_{11} = \log n + \log\log n \geq \log n + \omega(1)$, which is feasible according to condition (2.28).
The ATTRSPARSE algorithm requires that

$$mp_{11} - \log n \geq \omega(1),$$

which is also strictly covered by our condition (2.28). To illustrate the strict inclusion, consider the case where

$$p_{11} = \frac{2\log n}{n+m}, \quad p_{10} = p_{01} = 0, \text{ and } \quad m = n.$$

Then we have $mp_{11} = \log n < \log n + \omega(1)$, which is *not* feasible using the ATTRSPARSE algorithm. However, for such choice of p_{11} and m, we have $(n+m)p_{11} = 2\log n \geq \log n + \omega(1)$, which is feasible according to condition (2.28).

Comparison between the converse results: Our converse result in Corollary 2 further improves the best-known converse result in [20]. To see this, here we first simplify the converse result from Theorem 6. Note that under the condition $m = O(n)$, we have $\log(n+m) = \log n + O(1)$. Therefore, the converse condition in Theorem 6 is equivalent to

$$(n+m)p_{11} \leq \log n + O(1), \quad \text{and} \quad m = O(n).$$

Then we can directly see that in the regime $m = O(n)$, our specialized converse condition $(n+m)p_{11} \leq \log n - \omega(1)$ implies the converse in Theorem 6. However, in the $m = \omega(n)$ regime, Theorem 6 does not provide a condition for converse, while our converse condition is still valid when $m = \omega(n)$.

2.2.3 Bipartite graph pair

In the bipartite graph pair model $\mathcal{G}(n, m, \boldsymbol{q})$, each graph contains two disjoint and independent sets of vertices, i.e., the user vertex set $\mathcal{V}_{\mathrm{u}}$ and the attribute vertex set $\mathcal{V}_{\mathrm{a}}$. Edges between the two sets of vertices are generated in the following way: for $e \in \mathcal{V}_{\mathrm{u}} \times \mathcal{V}_{\mathrm{a}}$

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } q_{11}, \\ (1,0) & \text{w.p. } q_{10}, \\ (0,1) & \text{w.p. } q_{01}, \\ (0,0) & \text{w.p. } q_{00}. \end{cases} \tag{2.30}$$

The anonymized graph G_2' is obtained by applying a random permutation Π^* only on the user vertex set of G_2. The task of aligning such correlated bipartite graphs can be viewed as a special case of the database alignment problem, where the entries of the database are assumed to be independently drawn from the same Bernoulli distribution. [6] provides the best-known information-theoretic limits. Here, we state in Theorem 8 their results of this special case.

Theorem 8 (Specialization of Theorems 1 and 2 in [6] to bipartite graphs)**.** *Consider the bipartite graph pair model $\mathcal{G}(n, m, \boldsymbol{q})$.*
Achievability:
If

$$-\frac{m}{2}\log\left(1-2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2\right) \geq \log n + \omega(1), \tag{2.31}$$

then there exists an estimator that achieves exact alignment w.h.p.
Converse:
If

$$-\frac{m}{2}\log\left(1-2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2\right) \leq (1-\Omega(1))\log n \tag{2.32}$$

then any estimator achieves exact alignment with probability $o(1)$.

To compare our results with the best-known database alignment information-theoretic limits, we specialize the attributed Erdős–Rényi pair model to the bipartite graph pair by removing all of the edges between user vertices, i.e., setting $p_{00} = 1$. Correspondingly, we obtain the following achievability and converse results on bipartite graph alignment.

Theorem 9 (Specialization to bipartite graph pair model)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ with $p_{00} = 1$.*
Achievability:
If

$$m(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2 \geq \log n + \omega(1), \tag{2.33}$$

then the MAP estimator achieves exact alignment w.h.p.
Converse:
If

$$mq_{11} \leq \log n - \omega(1), \tag{2.34}$$

then no algorithm guarantees exact alignment w.h.p.

Remark 2. The achievability result from [3] (2.31) contains our specialized achievability result (2.33). To see this, recall that $\log(1+x) \leq x$ for all $x > -1$. Therefore we have $-\log\left(1-2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2\right) \geq 2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2$, and condition (2.33) implies condition (2.31).

Remark 3. Our specialized converse result (2.34) can slightly expand the best-known converse regime in (2.32). To give an example of the newly covered region, consider the case where $q_{11} = qs^2$, $q_{01} = q_{10} = qs(1-s)$ and $q_{00} = 1 - q_{11} - q_{01} - q_{10}$ for some $q = o(1)$ and $s = \Theta(1)$. Here we show that in this regime (2.32) implies (2.34). By our assumptions on $\boldsymbol{q}$, we have $(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2 = (1-o(1))qs^2 = o(1)$. For (2.32) we have that $-\frac{m}{2}\log(1-2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2) = mqs^2 - o(mqs^2) + q^2s^4(1-o(1)) \leq (1-\Omega(1))\log n$. Thus condition (2.32) implies that $mqs^2 \leq (1-\Omega(1))\log n + o(mqs^2)$. Here we also get that $mqs^2 = O(\log n)$ and then we have $o(mqs^2) = o(\log n)$. Finally, from condition (2.32), we can derive $mqs^2 \leq (1-\Omega(1))\log n + o(mqs^2) \leq \log n - \omega(1)$ which is exactly (2.34).

Chapter 3

Proof of the Achievability

In this chapter, we give detailed proof of the general achievability result (Theorem 1) and the achievability in the sparse region (Theorem 2). For both results, we prove by showing that the error probability of the MAP estimator converges to 0 as the number of users n goes to infinity. For Theorem 2, we focus on a regime where the edge density is relatively small, which allows us to obtain a tighter upper bound on the error event, and thus enlarge the achievable region from Theorem 1.

3.1 General Achievability (Theorem 1)

Here in this proof of the general achievability result, we begin by solving the exact alignment as an inference problem. To be more specific, the observed graph pair (G_1, G_2') are aligned using the MAP estimator, which further simplifies to a minimum weighted distance estimator (see Lemma 2). This minimum weighted distance estimator is optimal in the sense that it achieves the minimum probability of error. Along this line, in Chapter 3.1.2 we prove our general achievability result by showing that under some conditions on the attributed graph pairs $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$, the error probability of the minimum weighted distance estimator goes to 0 as n goes to infinity. To elaborate on this, in Chapter 3.1.4 we obtain a closed-form upper bound of this error probability using tools from enumerative combinatorics in Lemma 3, which are inspired by [4, 5]. From the obtained upper bound of error probability, we finally prove the general achievability result by figuring out sufficient conditions under which the closed-form upper bounds converge to 0 as n goes to infinity.

3.1.1 MAP estimation

In this section, we state a simple algorithm derived from MAP estimator, which is optimal for aligning graphs generated from $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$, and we defer the proof to Appendix A. Here we first introduce some basic notation for graph statistics needed in stating the MAP estimator. For any attributed

g on the vertex set $\mathcal{V}_{\mathrm{u}} \cup \mathcal{V}_{\mathrm{a}}$ and any permutation π over the user vertex set $\mathcal{V}_{\mathrm{u}}$, we use $\pi(g)$ to denoted the graph given by applying π to g. For any two attributed graphs g_1 and g_2 on $\mathcal{V}_{\mathrm{u}} \cup \mathcal{V}_{\mathrm{a}}$, we consider the Hamming distance between their edges restricted to the user-user vertex pairs in $\mathcal{E}_{\mathrm{u}}$, denoted as

$$\Delta^{\mathrm{u}}(g_1, g_2) = \sum_{(i,j)\in\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{g_1((i,j)) \neq g_2((i,j))\}; \tag{3.1}$$

and the Hamming distance between their edges restricted to the user-attribute vertex pairs in $\mathcal{E}_{\mathrm{a}}$, denoted as

$$\Delta^{\mathrm{a}}(g_1, g_2) = \sum_{(i,v)\in\mathcal{E}_{\mathrm{a}}} \mathbb{1}\{g_1((i,v)) \neq g_2((i,v))\}. \tag{3.2}$$

Lemma 2 (MAP estimator)**.** *Let (G_1, G_2') be an observable pair generated from the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. The MAP estimator of the permutation Π^* based on (G_1, G_2') simplifies to*

$$\begin{aligned}&\hat{\pi}_{\mathrm{MAP}}(G_1, G_2')\\ &= \operatorname*{argmin}_{\pi\in\mathcal{S}_n}\{w_1\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + w_2\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))\},\end{aligned}$$

where $w_1 = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)$, $w_2 = \log\left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)$, *and*

$$\begin{aligned}\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) &= \sum_{(i,j)\in\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1((i,j)) \neq G_2'((\pi(i),\pi(j)))\},\\ \Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2')) &= \sum_{(i,v)\in\mathcal{E}_{\mathrm{a}}} \mathbb{1}\{G_1((i,v)) \neq G_2'((\pi(i),v))\}.\end{aligned}$$

3.1.2 Proof of the general achievability (Theorem 1)

Theorem 1 (General achievability)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If*

$$\tfrac{n\psi_{\mathrm{u}}}{2} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{2.1}$$

then the MAP estimator achieves exact alignment w.h.p.

Proof of Theorem 1. Given the observable pair (G_1, G_2'), the error probability of MAP estimator can be upper-bounded as

$$\mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') \neq \Pi^*)$$

$$
\begin{aligned}
&= \sum_{\pi^* \in \mathcal{S}_n} \mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') \neq \pi^* | \Pi^* = \pi^*) \mathsf{P}(\Pi^* = \pi^*) \\
&= \frac{1}{|\mathcal{S}_n|} \sum_{\pi^* \in \mathcal{S}_n} \mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') \neq \pi^* | \Pi^* = \pi^*) && (3.3) \\
&= \mathsf{P}\left(\hat{\pi}_{\mathrm{MAP}}(G_1, G_2) \neq \pi_{\mathrm{id}} | \Pi^* = \pi_{\mathrm{id}}\right) && (3.4) \\
&= \mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1, G_2) \neq \pi_{\mathrm{id}}) && (3.5) \\
&\leq \mathsf{P}(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0), && (3.6)
\end{aligned}
$$

where π_{id} denotes the identity permutation, and

$$
\begin{aligned}
\delta_\pi(G_1, G_2) \triangleq\ & w_1(\Delta^{\mathrm{u}}(G_1, \pi(G_2)) - \Delta^{\mathrm{u}}(G_1, G_2)) \\
& + w_2(\Delta^{\mathrm{a}}(G_1, \pi(G_2)) - \Delta^{\mathrm{a}}(G_1, G_2)). \quad (3.7)
\end{aligned}
$$

Here (3.3) follows from the fact that Π^* is uniformly drawn from $\mathcal{S}_n$, which implies $\mathsf{P}(\Pi^* = \pi^*) = 1/|\mathcal{S}_n|$ for all π^*; (3.4) is due to the symmetry among user vertices in G_1 and G_2; (3.5) is due to the independence between Π^* and (G_1, G_2); (3.6) is true because by Lemma 2, $\pi_{\mathrm{MAP}}(G_1, G_2)$ minimizes the weighted distance, and $\pi_{\mathrm{MAP}} \neq \pi_{\mathrm{id}}$ only if there exists a permutation π such that $\pi \neq \pi_{\mathrm{id}}$ and $\delta_\pi(G_1, G_2) \leq 0$.

Now to prove that (2.1) implies that the error probability in (3.6) converges to 0 as $n \to \infty$, we further upper-bound the error probability as follows

$$
\begin{aligned}
&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0\right) \\
&\leq \sum_{\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}} \mathsf{P}(\delta_\pi(G_1, G_2) \leq 0) && (3.8) \\
&= \sum_{\tilde{n}=2}^{n} \sum_{\pi \in \mathcal{S}_{n,\tilde{n}}} \mathsf{P}(\delta_\pi(G_1, G_2) \leq 0) && (3.9) \\
&\leq \sum_{\tilde{n}=2}^{n} |\mathcal{S}_{n,\tilde{n}}| \max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)\} \\
&\leq \sum_{\tilde{n}=2}^{n} n^{\tilde{n}} \max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)\}.
\end{aligned}
$$

Here (3.8) follows from directly applying the union bound. In (3.9), we use $\mathcal{S}_{n,\tilde{n}}$ to denote the set of permutations on $[n]$ that contains exactly $(n - \tilde{n})$ fixed points. In the example of Figure 1.1, the given permutation $\Pi^* = (1)(23)$ has 1 fixed point and $(1)(23) \in \mathcal{S}_{3,2}$. Furthermore, we have

$|\mathcal{S}_{n,\tilde{n}}| = \binom{n}{\tilde{n}} (!\tilde{n}) \leq n^{\tilde{n}}$, where $!\tilde{n}$, known as the number of derangements, represents the number of permutations on a set of size $\tilde{n}$ such that no element appears in its original position.

Now, we apply Lemma 3 to obtain a closed-form upper bound on the term $\max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)\}$. Lemma 3 is stated after this proof and its proof is presented in Chapter 3.1.4. With the upper bound in Lemma 3, we have

$$\begin{aligned}
&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0\right) \\
&\leq \sum_{\tilde{n}=2}^{n} n^{\tilde{n}} (1-2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}} (1-2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}} \\
&= \sum_{\tilde{n}=2}^{n} \left(n(1-2\psi_{\mathrm{u}})^{\frac{n-2}{4}} (1-2\psi_{\mathrm{a}})^{\frac{m}{2}} \right)^{\tilde{n}}. \qquad (3.10)
\end{aligned}$$

For this geometry series, the negative logarithm of its common ratio is

$$\begin{aligned}
&-\log n - \frac{n-2}{4} \log(1-2\psi_{\mathrm{u}}) - \frac{m}{2} \log(1-2\psi_{\mathrm{a}}) \\
&\geq -\log n + \frac{n-2}{2} \psi_u + m\psi_a \qquad (3.11) \\
&= \omega(1). \qquad (3.12)
\end{aligned}$$

Here we have $\psi_{\mathrm{u}} = (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \leq 1/4$ and $\psi_{\mathrm{u}} = (\sqrt{q_{11}q_{00}} - \sqrt{q_{10}q_{01}})^2 \leq 1/4$. Therefore, (3.11) follows from the inequality $\log(1+x) \leq x$ for $x > -1$. Equation (3.12) follows from condition (2.1) by noting that ψ_u is no larger than 1. Therefore, the geometry series in (3.10) converge to 0 as $n \to \infty$. This completes the proof that MAP estimator achieves exact alignment w.h.p. under condition (2.1). □

Lemma 3. *Let (G_1, G_2) be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. For any permutation π, let*

$$\begin{aligned}
\delta_\pi(G_1, G_2) \triangleq\ & w_1(\Delta^{\mathrm{u}}(G_1, \pi(G_2)) - \Delta^{\mathrm{u}}(G_1, G_2)) \\
& + w_2(\Delta^{\mathrm{a}}(G_1, \pi(G_2)) - \Delta^{\mathrm{a}}(G_1, G_2)).
\end{aligned}$$

Then when π has $n - \tilde{n}$ fixed points, we have

$$\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0\right) \leq (1-2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}} (1-2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}}.$$

3.1.3 An interlude of generating functions

To prove the upper bound on $\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)$ in Lemma 3, we will use the method of *generating functions*. In this section, we first introduce our construction of a generating function and how it can be used to bound $\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)$. We then present several properties of generating functions (Facts 1, 2, and 3), which will be needed in the proof of Lemma 3.

A generating function for the attributed Erdős–Rényi pair. For any graph pair (g, h) that is a realization of an attributed Erdős–Rényi pair, we define a 2×2 matrix $\boldsymbol{\mu}(g, h)$ as follows for user-user edges:

$$\boldsymbol{\mu}(g, h) = \begin{pmatrix} \mu_{11} & \mu_{10} \\ \mu_{01} & \mu_{00} \end{pmatrix},$$

where $\mu_{ij} = \sum_{e \in \mathcal{E}_{\mathrm{u}}} \mathbb{1}\{g(e) = i, h(e) = j\}$. Similarly, we define $\boldsymbol{\nu}(g, h)$ as follows for user-attribute edges:

$$\boldsymbol{\nu}(g, h) = \begin{pmatrix} \nu_{11} & \nu_{10} \\ \nu_{01} & \nu_{00} \end{pmatrix},$$

where $\nu_{ij} = \sum_{e \in \mathcal{E}_{\mathrm{a}}} \mathbb{1}\{g(e) = i, h(e) = j\}$.

Now we define a generating function for attributed graph pairs, which encodes information in a formal power series. Let z be a single formal variable and $\boldsymbol{x}$ and $\boldsymbol{y}$ be 2×2 matrices of formal variables where

$$\boldsymbol{x} = \begin{pmatrix} x_{00} & x_{01} \\ x_{10} & x_{11} \end{pmatrix} \text{ and } \boldsymbol{y} = \begin{pmatrix} y_{00} & y_{01} \\ y_{10} & y_{11} \end{pmatrix}.$$

Then for each permutation π, we construct the following generating function:

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) = \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}, \tag{3.13}$$

where

$$\boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \triangleq x_{00}^{\mu_{00}} \cdot x_{01}^{\mu_{01}} \cdot x_{10}^{\mu_{10}} \cdot x_{11}^{\mu_{11}},$$
$$\boldsymbol{y}^{\boldsymbol{\nu}(g,h)} \triangleq y_{00}^{\nu_{00}} \cdot y_{01}^{\nu_{01}} \cdot y_{10}^{\nu_{10}} \cdot y_{11}^{\nu_{11}}.$$

Note that in the above expression of $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$, we enumerate all possible attributed graph pairs (g, h) as realizations of the random graph pair (G_1, G_2). For each realization, we encode the corresponding $\boldsymbol{\mu}(g, h), \boldsymbol{\nu}(g, h)$ and $\delta_\pi(g, h)$ in the powers of formal variables $\boldsymbol{x}, \boldsymbol{y}$ and z. By summing over all possible realizations (g, h), the terms having the same powers are merged as one term. Therefore, the coefficient of a term $z^{\delta_\pi} \boldsymbol{x}^{\boldsymbol{\mu}} \boldsymbol{y}^{\boldsymbol{\nu}}$ represents the

number of graph pairs that have the same graph statics represented in the powers of formal variables.

Bounding $\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0)$ ***in terms of the generating function.*** We first argue that when we set $\boldsymbol{x} = \boldsymbol{p}$ and $\boldsymbol{y} = \boldsymbol{q}$, the generating function $\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z)$ becomes the probability generating function of $\delta_\pi(G_1, G_2)$ for the attributed Erdős–Rényi pair $(G_1, G_2) \sim \mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. To see this, note that the joint distribution of G_1 and G_2 can be written as $\mathsf{P}((G_1, G_2) = (g, h)) = \boldsymbol{p}^{\boldsymbol{\mu}(g,h)}\boldsymbol{q}^{\boldsymbol{\nu}(g,h)}$. Then by combining terms in $\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z)$, we have $\mathsf{P}(\delta_\pi(G_1, G_2) = d) = [z^d]\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z)$, where $[z^d]\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z)$ denotes the coefficient of z^d with $[z^d]$ being the *coefficient extraction operator*. We comment that the probability generating function here is defined in the sense that $\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z) = \mathsf{E}\big[z^{\delta_\pi(G_1,G_2)}\big]$. Since $\delta_\pi(G_1, G_2)$ takes real values, this is slightly different from the standard probability generating function for random variables with nonnegative integer values. But this distinction does not affect our analysis in a significant way since $\delta_\pi(G_1, G_2)$ takes values from a finite set.

Now it is easy to see that

$$\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0) = \sum_{d \leq 0} [z^d]\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z). \tag{3.14}$$

Cycle decomposition. We will use the cycle decomposition of permutations to simply the form of the generating function $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$.

Each permutation π induces a permutation on the vertex-pair set. We denote this induced permutation as $\pi^{\mathcal{E}}$, where $\pi^{\mathcal{E}} : \mathcal{E} \to \mathcal{E}$ and $\pi^{\mathcal{E}}((u, v)) = (\pi(u), \pi(v))$ for $u, v \in \mathcal{V}$. A *cycle* of the induced permutation $\pi^{\mathcal{E}}$ is a list of vertex pairs such that each vertex pair is mapped to the vertex pair next to it in the list (with the last mapped to the first one). The cycles naturally partition the set of vertex pairs, $\mathcal{E}$, into disjoint subsets where each subset consists of the vertex pairs from a cycle. We refer to each of these subsets as an *orbit*. For the example given in Figure 1.1, the induced permutation on $\mathcal{E}$ can divide it into 4 orbits of size 1 (1-orbit): $\{(2, 3)\}$, $\{(1, a)\}$, $\{(1, b)\}$, $\{(1, c)\}$, and 4 orbits of length 2 (2-orbit): $\{(1, 2), (1, 3)\}$, $\{(2, a), (3, a)\}$, $\{(2, b), (3, b)\}$, $\{(2, c), (3, c)\}$.

We write this partition of $\mathcal{E}$ based on the cycle decomposition as $\mathcal{E} = \cup_{k \geq 1}\mathcal{O}_k$, where $\mathcal{O}_k$ denotes the kth orbit. Note that each cycle consists of either only user-user vertex pairs or only user-attribute vertex pairs. If a single orbit $\mathcal{O}_k$ contains only user-user vertex pairs, we define its generating function on formal variables z and $\boldsymbol{x}$ as

$$\mathcal{A}_{\mathcal{O}_k}(\boldsymbol{x}, z) = \sum_{g \in \{0,1\}^{\mathcal{O}_k}} \sum_{h \in \{0,1\}^{\mathcal{O}_k}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)}.$$

If $\mathcal{O}_k$ contains only user-attribute vertex pairs, we define its generating function on formal variables z and $\boldsymbol{y}$ as

$$\mathcal{A}_{\mathcal{O}_k}(\boldsymbol{y}, z) = \sum_{g\in\{0,1\}^{\mathcal{O}_k}} \sum_{h\in\{0,1\}^{\mathcal{O}_k}} z^{\delta_\pi(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}.$$

Here, we extend the previous definitions of δ_π, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ on attributed graphs to any set of vertex pairs. Let $\mathcal{E}'$ be an arbitrary set of vertex pairs. Then we define δ_π for any $g, h \in \{0,1\}^{\mathcal{E}'}$ as

$$\begin{aligned}\delta_\pi(g,h) =& w_1 \sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{u}}} \left(\mathbb{1}\{g(e) \neq h(\pi^{\mathcal{E}}(e))\} - \mathbb{1}\{g(e)\neq h(e)\}\right) \\ &+ w_2 \sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{a}}} \left(\mathbb{1}\{g(e) \neq h(\pi^{\mathcal{E}}(e))\} - \mathbb{1}\{g(e)\neq h(e)\}\right). \end{aligned} \tag{3.15}$$

For $g, h \in \{0,1\}^{\mathcal{E}'}$, we keep $\boldsymbol{\mu}(g,h)$ and $\boldsymbol{\nu}(g,h)$ as 2×2 matrices as follows:

$$\boldsymbol{\mu}(g,h) = \begin{pmatrix}\mu_{11}, \mu_{10}\\ \mu_{01}, \mu_{00}\end{pmatrix} \quad \text{and} \quad \boldsymbol{\nu}(g,h) = \begin{pmatrix}\nu_{11}, \nu_{10}\\ \nu_{01}, \nu_{00}\end{pmatrix},$$

where

$$\mu_{ij} = \mu_{ij}(g,h) \triangleq \sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{g(e)=i, h(e)=j\}, \tag{3.16}$$

$$\nu_{ij} = \nu_{ij}(g,h) \triangleq \sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{a}}} \mathbb{1}\{g(e)=i, h(e)=j\}. \tag{3.17}$$

We remind the reader that by setting the set of vertex pairs $\mathcal{E}'$ to be $\mathcal{E}$ these extended definitions on δ_π, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ agree with the previous definition where g, h are attributed graphs.

Now, we consider the generating functions for two orbits $\mathcal{O}_k$ and $\mathcal{O}_{k'}$. If the size of $\mathcal{O}_k$ equals to the size of $\mathcal{O}_{k'}$ and both orbits consist of user-user vertex pairs, then we claim that $\mathcal{A}_{\mathcal{O}_k}(\boldsymbol{x}, z) = \mathcal{A}_{\mathcal{O}_{k'}}(\boldsymbol{x}, z)$. This is because to obtain $\mathcal{A}_{\mathcal{O}_k}(\boldsymbol{x}, z)$, we sum over all realizations $g, h \in \{0,1\}^{\mathcal{O}_k}$, which is equivalent to summing over $g, h \in \{0,1\}^{\mathcal{O}_{k'}}$. Similarly, if the size of $\mathcal{O}_k$ equals to the size of $\mathcal{O}_{k'}$ and both orbits consist of user-attribute vertex pairs, we have $\mathcal{A}_{\mathcal{O}_k}(\boldsymbol{y}, z) = \mathcal{A}_{\mathcal{O}_{k'}}(\boldsymbol{y}, z)$. To make the notation compact, we define a generating function $\mathcal{A}_l(\boldsymbol{x}, z)$ for size l user-user orbits and a generating function $\mathcal{A}_l(\boldsymbol{y}, z)$ for size l user-attribute orbits. Let $\mathcal{E}_l^{\mathrm{u}}$ denote a general user-user orbit of size l and $\mathcal{E}_l^{\mathrm{a}}$ denote a general user-attribute orbit of size l. Then

$$\mathcal{A}_l(\boldsymbol{x}, z) \triangleq \sum_{g\in\{0,1\}^{\mathcal{E}_l^{\mathrm{u}}}} \sum_{h\in\{0,1\}^{\mathcal{E}_l^{\mathrm{u}}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)}, \tag{3.18}$$

$$\mathcal{A}_l(\boldsymbol{y}, z) \triangleq \sum_{g \in \{0,1\}^{\mathcal{E}_l^{\mathrm{a}}}} \sum_{h \in \{0,1\}^{\mathcal{E}_l^{\mathrm{a}}}} z^{\delta_\pi(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}. \tag{3.19}$$

Properties of generating functions. Here we provide some properties of the generating functions defined above. These properties are mainly used for upper bounding error probability of the MAP estimator through probability generating functions.

Fact 1. *The generating function $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$ of permutation π can be decomposed into*

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) = \prod_{l \geq 1} \mathcal{A}_l(\boldsymbol{x}, z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{y}, z)^{t_l^{\mathrm{a}}},$$

where t_l^{u} is the number of user-user orbits of size l, t_l^{a} is the number of user-attribute orbits of size l.

Fact 2. *Let $\boldsymbol{x} \in \mathbb{R}^{2\times 2}$ and $z \neq 0$. Then for all $l \geq 2$, we have $\mathcal{A}_l(\boldsymbol{x}, z) \leq \mathcal{A}_2(\boldsymbol{x}, z)^{\frac{l}{2}}$ and $\mathcal{A}_l(\boldsymbol{x}, z) \leq \mathcal{A}_2(\boldsymbol{x}, z)^{\frac{l}{2}}$.*

We refer the readers to Appendix C for the proof of Fact 1, and Theorem 4 in [5] for the proof of Fact 2. Combining these two facts, we get

$$\begin{aligned}\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) \leq & \mathcal{A}_1(\boldsymbol{x}, z)^{t_1^{\mathrm{u}}} \mathcal{A}_1(\boldsymbol{y}, z)^{t_1^{\mathrm{a}}} \\ & \mathcal{A}_2(\boldsymbol{x}, z)^{\frac{t^{\mathrm{u}} - t_1^{\mathrm{u}}}{2}} \mathcal{A}_2(\boldsymbol{x}, z)^{\frac{nm - t_1^{\mathrm{a}}}{2}}.\end{aligned} \tag{3.20}$$

Here, in (3.20), we use t^{u} to denote the total number of user-user pairs and $t^{\mathrm{u}} = \sum_{l \geq 1} t_l^{\mathrm{u}} l = \binom{n}{2}$. We have the closed-form expressions for $\mathcal{A}_1$ and $\mathcal{A}_2$ following from their definition in (3.18) and (3.19)

$$\mathcal{A}_1(\boldsymbol{x}, z) = x_{00} + x_{10} + x_{01} + x_{11}, \tag{3.21}$$

$$\mathcal{A}_1(\boldsymbol{y}, z) = y_{00} + y_{10} + y_{01} + y_{11}, \tag{3.22}$$

$$\begin{aligned}\mathcal{A}_2(\boldsymbol{x}, z) = & (x_{00} + x_{10} + x_{01} + x_{11})^2 \\ & + 2x_{00}x_{11}(z^{2w_1} - 1) + 2x_{10}x_{01}(z^{-2w_1} - 1),\end{aligned} \tag{3.23}$$

$$\begin{aligned}\mathcal{A}_2(\boldsymbol{y}, z) = & (y_{00} + y_{10} + y_{01} + y_{11})^2 \\ & + 2y_{00}y_{11}(z^{2w_2} - 1) + 2y_{10}y_{01}(z^{-2w_2} - 1).\end{aligned} \tag{3.24}$$

Moreover, we have Fact 3 which gives explicit upper bounds on the coefficients of a generating function

Fact 3. *For a discrete random variable X defined over a finite set $\mathcal{X}$, let*

$$\Phi(z) \triangleq \mathsf{E}[z^X] = \sum_{i \in \mathcal{X}} \mathsf{P}(X = i) z^i \tag{3.25}$$

be the probability generating function of X. Then, for any $j \in \mathcal{X}$ and $z > 0$,

$$[z^j]\Phi(z) \leq z^{-j}\Phi(z). \tag{3.26}$$

For any $j \in \mathcal{X}$ and $z \in (0, 1]$,

$$\sum_{\substack{i \leq j \\ i \in \mathcal{X}}} [z^i]\Phi(z) \leq z^{-j}\Phi(z). \tag{3.27}$$

For any $j \in \mathcal{X}$ and $z \geq 1$,

$$\sum_{\substack{i \geq j \\ i \in \mathcal{X}}} [z^i]\Phi(z) \leq z^{-j}\Phi(z). \tag{3.28}$$

Proof. We write $p_i \triangleq \mathsf{P}(X = i)$ in this proof. For any $j \in \mathcal{X}$ and $z > 0$, we have

$$z^{-j}\Phi(z) - [z^j]\Phi(z) = \sum_{i \in \mathcal{X}} p_i z^{i-j} - p_j = \sum_{\substack{i \neq j \\ i \in \mathcal{X}}} p_i z^{i-j} \geq 0,$$

which establishes (3.26).

For any $j \in \mathcal{X}$ and $z \in (0, 1)$, we have $\sum_{i \leq j} p_i \leq \sum_{i \leq j} p_i z^{i-j}$. Therefore, we have

$$\sum_{\substack{i \leq j \\ i \in \mathcal{X}}} [z^i]\Phi(z) = \sum_{\substack{i \leq j \\ i \in \mathcal{X}}} p_i \leq \sum_{\substack{i \leq j \\ i \in \mathcal{X}}} p_i z^{i-j} \leq \sum_{i \in \mathcal{X}} p_i z^{i-j} = z^{-j}\Phi(z),$$

which establishes (3.27).

For any $z > 1$ and $j \in \mathcal{X}$, we have $\sum_{i \geq j} p_i \leq \sum_{i \geq j} p_i z^{i-j}$. Therefore, we have

$$\sum_{\substack{i \geq j \\ i \in \mathcal{X}}} [z^i]\Phi(z) = \sum_{\substack{i \geq j \\ i \in \mathcal{X}}} p_i \leq \sum_{\substack{i \geq j \\ i \in \mathcal{X}}} p_i z^{i-j} \leq \sum_{i \in \mathcal{X}} p_i z^{i-j} = z^{-j}\Phi(z),$$

which establishes (3.28). □

3.1.4 Upper bound on the error event of MAP (Lemma 3)

Now with techniques about generating functions from last section, we are ready to prove Lemma 3.

Lemma 3. *Let (G_1, G_2) be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. For any permutation π, let*

$$\delta_\pi(G_1, G_2) \triangleq w_1(\Delta^{\mathrm{u}}(G_1, \pi(G_2)) - \Delta^{\mathrm{u}}(G_1, G_2)) + w_2(\Delta^{\mathrm{a}}(G_1, \pi(G_2)) - \Delta^{\mathrm{a}}(G_1, G_2)).$$

Then when π has $n - \tilde{n}$ fixed points, we have

$$\mathsf{P}\left(\delta_\pi(G_1, G_2) \le 0\right) \le (1 - 2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}} (1 - 2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}}.$$

Proof. For any $\pi \in \mathcal{S}_{n,\tilde{n}}$ and any $z_1 \in (0, 1)$, we have

$$\begin{aligned}
&\mathsf{P}\left(\delta_\pi(G_1, G_2) \le 0\right) \\
&= \sum_{d \le 0} [z^d] \mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z) \\
&\le \mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z_1) && (3.29) \\
&\le \mathcal{A}_1(\boldsymbol{p}, z)^{t_1^{\mathrm{u}}} \mathcal{A}_1(\boldsymbol{q}, z)^{t_1^{\mathrm{a}}} \\
&\quad \mathcal{A}_2(\boldsymbol{p}, z)^{\frac{t^{\mathrm{u}} - t_1^{\mathrm{u}}}{2}} \mathcal{A}_2(\boldsymbol{q}, z)^{\frac{nm - t_1^{\mathrm{a}}}{2}} && (3.30) \\
&\le \mathcal{A}_2(\boldsymbol{p}, z)^{\frac{t^{\mathrm{u}} - t_1^{\mathrm{u}}}{2}} \mathcal{A}_2(\boldsymbol{q}, z)^{\frac{nm - t_1^{\mathrm{a}}}{2}}. && (3.31)
\end{aligned}$$

In (3.29), we set $z \in (0, 1)$, and this upper bound follows from Fact 3. (3.30) follows from the decomposition on $\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z)$ stated in Fact 1. Equation (3.31) follows since $\mathcal{A}_1(\boldsymbol{p}, z) = \mathcal{A}_1(\boldsymbol{q}, z) = 1$ according to their expression in (3.21) and (3.22). To obtain a tight bound, we then search for $z \in (0, 1)$ that achieves the minimum of (3.31). Following the definition of $\mathcal{A}_2(\boldsymbol{p}, z)$ in (3.23) and using the inequality $a/x + bx \ge 2\sqrt{ab}$, we have

$$\begin{aligned}
\mathcal{A}_2(\boldsymbol{p}, z) &= 1 + 2p_{00}p_{11}(z^{2w_1} - 1) + 2p_{10}p_{01}(z^{-2w_1} - 1) \\
&\ge 1 - 2p_{00}p_{11} - 2p_{10}p_{01} + 4\sqrt{p_{00}p_{11}p_{10}p_{01}} \\
&= 1 - 2(\sqrt{p_{00}p_{11}} - \sqrt{p_{10}p_{01}})^2 \triangleq 1 - 2\psi_{\mathrm{u}}. && (3.32)
\end{aligned}$$

Here the equality holds if and only if $z^{2w_1} = \sqrt{\frac{p_{10}p_{01}}{p_{00}p_{11}}}$. Recall that $w_1 = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)$. Therefore, $\mathcal{A}_1(\boldsymbol{p}, z)$ achieves the minimum when $z = e^{-1/4}$. Similarly, we have

$$\mathcal{A}_2(\boldsymbol{q}, z) = 1 + 2q_{00}q_{11}(z^{2w_2} - 1) + 2q_{10}q_{01}(z^{-2w_2} - 1)$$

$$\geq 1 - 2q_{00}q_{11} - 2q_{10}q_{01} + 4\sqrt{q_{00}q_{11}q_{10}q_{01}}$$
$$= 1 - 2(\sqrt{q_{00}q_{11}} - \sqrt{q_{10}q_{01}})^2 \triangleq 1 - 2\psi_{\mathrm{a}}. \tag{3.33}$$

Here the equality holds if and only if $z^{2w_2} = \sqrt{\frac{q_{10}q_{01}}{q_{00}q_{11}}}$. With $w_2 = \log\left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)$, we have that $\mathcal{A}_2(\boldsymbol{q}, z)$ achieves the minimum when $z = e^{-1/4}$. Therefore, $z = e^{-1/4}$ minimizes (3.31) and we have

$$\begin{aligned}
&\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0\right) \\
&\leq (1 - 2\psi_{\mathrm{u}})^{\frac{t^{\mathrm{u}} - t_1^{\mathrm{u}}}{2}} (1 - 2\psi_{\mathrm{a}})^{\frac{mn - t_1^{\mathrm{a}}}{2}} \\
&\leq (1 - 2\psi_{\mathrm{u}})^{\frac{\tilde{n}(2n - \tilde{n} - 2)}{4}} (1 - 2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}} \qquad (3.34) \\
&\leq (1 - 2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}} (1 - 2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}}. \qquad (3.35)
\end{aligned}$$

In (3.34), we use the following relations between the number of fixed vertex pairs t_1^{u}, t_1^{a} and number of fixed vertices $\tilde{n}$

$$\begin{aligned}
&\binom{n - \tilde{n}}{2} \leq t_1^{\mathrm{u}} \leq \binom{n - \tilde{n}}{2} + \frac{\tilde{n}}{2}, \qquad (3.36) \\
&t_1^{\mathrm{a}} = (n - \tilde{n})m.
\end{aligned}$$

In the given upper bound of t_1^{u}, $\binom{n-\tilde{n}}{2}$ corresponds to the number of user-user vertex pairs whose two vertices are both fixed under π, and $\frac{\tilde{n}}{2}$ is the upper bound of user-user vertex pairs whose two vertices are swapped under π. In (3.35), we use the fact that $\tilde{n} \leq n$. □

3.2 Achievability in Sparse Region (Theorem 2)

In this section, we prove Theorem 2, which characterizes the achievable region when the user-user connection is *sparse* in the sense that $p_{11} = O(\frac{\log n}{n})$. We use R to denote the number of user-user edges in the intersection graph and it follows a binomial distribution $\mathrm{Bin}(t^{\mathrm{u}}, p_{11})$. In the sparse regime where $p_{11} = O(\frac{\log n}{n})$, the achievability proof here is different from what we did in Chapter 3.1. The reason for applying a different proof technique is that, in this sparse regime, the union bound we applied in Chapter 3.1 on $\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0\right)$ becomes very loose. To elaborate on this point, notice that the error of union bound comes from counting the intersection events multiple times. Therefore, if the probability of such intersection events becomes larger, then the union bound will be looser. In our problem, our event space contains sets of possible realizations on (G_1, G_2)

and an example of the aforementioned intersection events is $\{R = 0\}$ which lays in the intersection of $\{\delta_\pi(G_1, G_2) \leq 0\}$ for all $\pi \in \mathcal{S}_n$. Moreover, other events where R is small are also in the intersection of $\{\delta_\pi(G_1, G_2) \leq 0\}$ for some $\pi \in \mathcal{S}_n$ and the number of such permutations (equivalently the times of repenting when apply union bound) increases as R gets smaller. As a result, if p_{11} becomes relatively small, then the probability that R is small will be large and thus union bound will be loose.

To overcome the problem of the loose union bound in the sparse regime, we apply a *truncated union bound*. We first expand the probability we want to bound as follows

$$\begin{aligned}&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0\right)\\ =&\sum_{r\geq 0} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 | R = r\right) \mathsf{P}(R = r).\end{aligned}$$

We then apply the union bound on the conditional probability on the error event $\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 | R = r\right)$. As we discussed before, the error of applying union bound directly should be a function on r. Therefore, for some small r, the union bound on $\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 | R = r\right)$ is very loose while for the other r, the union bound is relatively tight. Therefore, we truncate the union bound on the conditional probability by taking the minimum with 1, which is an upper bound for any probability

$$\begin{aligned}&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 | R = r\right)\\ &= \min\{1, \sum_{\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}} \mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 | R = r\right)\}.\end{aligned}$$

Through such truncating, we avoid adopting the union bound when it is too loose and obtain a tighter bound. For example, given that $R = 0$, we have $\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 | R = 0\right) = 1$ for all $\pi \in \mathcal{S}_n$. Thus, by using the truncated union bound, we obtain 1 as a the upper bound instead of $(n! - 1)$. Overall, the key idea of our proof is first derive $\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 | R = r\right)$ as a function of r and then apply the truncated union bound according to how large this conditional probability is. This idea is inspired by [5] and is extended to the attributed Erdős–Rényi pair model. We restate the lemma to prove as follows.

Theorem 2 (Achievability in sparse region)**.** *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If*

$$p_{11} = O\left(\frac{\log n}{n}\right), \tag{2.2}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{2.3}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{2.4}$$

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{2.5}$$

then the MAP estimator achieves exact alignment w.h.p.

Proof of Theorem 2. We discuss two regimes $p_{11} = O(\frac{1}{n})$ and $\omega(\frac{1}{n}) \le p_{11} \le \Theta(\frac{\log n}{n})$.

When $p_{11} = O(\frac{1}{n})$, we have $n\psi_{\mathrm{u}} \le np_{11} = O(1)$. Thus, the sufficient condition (2.1) for exact alignment in Theorem 1

$$\tfrac{n\psi_{\mathrm{u}}}{2} + m\psi_{\mathrm{a}} - \log n = \omega(1)$$

is satisfied when condition (2.5)

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1)$$

is satisfied. By Theorem 1 , exact alignment is achievable w.h.p.

Now suppose $\omega(\frac{1}{n}) \le p_{11} \le \Theta(\frac{\log n}{n})$. Note that the number of edges in the intersection graph of G_1 and G_2 has the following equivalent representation

$$R = \mu_{11}(G_1, G_2) = \sum_{e \in \mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1(e) = 1, G_2(e) = 1\}.$$

Then, $R \sim \mathrm{Bin}(t^{\mathrm{u}}, p_{11})$ and $\mathsf{E}[R] = t^{\mathrm{u}}p_{11} = \binom{n}{2}p_{11} = \omega(n)$. By the Chebyshev's inequality, for any constant $\epsilon > 0$,

$$\mathsf{P}(|R - \mathsf{E}[R]| \ge \epsilon \mathsf{E}[R]) \le \frac{\mathsf{Var}(R)}{\epsilon^2 \mathsf{E}[R]^2} = \frac{1 - p_{11}}{\epsilon^2} \frac{1}{\mathsf{E}[R]} = o\left(\frac{1}{n}\right).$$

In the following, we upper bound the probability of error by discussing two cases: when $R \le (1+\epsilon)\mathsf{E}[R]$ and when $R > (1+\epsilon)\mathsf{E}[R]$. We have

$$\begin{aligned}
&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0\right) \\
&= \sum_{r=0}^{t^{\mathrm{u}}} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \backslash \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) \\
&\le \sum_{r \le (1+\epsilon)\mathsf{E}[R]} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \backslash \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) + \mathsf{P}\left(|R - \mathsf{E}[R]| > \epsilon \mathsf{E}[R]\right) \\
&= \sum_{r \le (1+\epsilon)\mathsf{E}[R]} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \backslash \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) + o(1) \qquad (3.37)
\end{aligned}$$

$$\leq \sum_{r\leq(1+\epsilon)\mathsf{E}[R]} 3n^2 z_6^2 \mathsf{P}(R=r) + o(1) \tag{3.38}$$

$$= \sum_{r\leq(1+\epsilon)\mathsf{E}[R]} 3n^2 z_6^2 \binom{t^{\mathrm{u}}}{r} p_{11}^r (1-p_{11})^{t^{\mathrm{u}}-r} + o(1)$$

$$= 3n^2(1-2\psi_{\mathrm{a}})^m \sum_{r=0}^{t^{\mathrm{u}}} \binom{t^{\mathrm{u}}}{r} p_{11}^r e^{-\frac{4r}{n}} (1-p_{11})^{t^{\mathrm{u}}-r} + o(1)$$

$$= 3n^2(1-2\psi_{\mathrm{a}})^m \left(p_{11}e^{-\frac{4}{n}} + 1 - p_{11}\right)^{t^{\mathrm{u}}} + o(1) \tag{3.39}$$

$$\leq 3n^2(1-2\psi_{\mathrm{a}})^m \left(1-\tfrac{4}{n}p_{11}\right)^{t^{\mathrm{u}}} + o(1). \tag{3.40}$$

Here (3.37) follows from the Chebyshev's inequality above. In (3.38), $z_6 = \exp\{-\frac{2r}{n} + \frac{m}{2}\log(1-2\psi_{\mathrm{a}}) + O(1)\}$. This step will be justified by Lemma 4, which is the major technical step in establishing the error bound. To apply Lemma 4, we need the conditions (2.2) (2.3) (2.4) and $r = O(\mathsf{E}[R]) = O(n\log n)$ to hold and we will explain the reason in the proof of Lemma 4. Equation (3.39) follows from the binomial formula and (3.40) follows from the inequality $e^x - 1 \leq x$. Taking the negative logarithm of the first term in (3.40), we have

$$\begin{aligned}
&-\log\left(3n^2(1-2\psi_{\mathrm{a}})^m \left(1-\tfrac{4}{n}p_{11}\right)^{t^{\mathrm{u}}}\right) \\
&= -2\log n - m\log(1-2\psi_{\mathrm{a}}) - t^{\mathrm{u}}\log\left(1-\tfrac{4p_{11}}{n}\right) + O(1) \\
&\geq -2\log n + 2m\psi_{\mathrm{a}} + t^{\mathrm{u}}\frac{4p_{11}}{n} + O(1) && (3.41)\\
&= -2\log n + 2m\psi_{\mathrm{a}} + 2np_{11} + O(1) && (3.42)\\
&= \omega(1). && (3.43)
\end{aligned}$$

Here, we have (3.41) follows from the inequality $\log(1+x) \leq x$ for $x > -1$. We get equation (3.42) by plugging in $t^{\mathrm{u}} = \binom{n}{2}$. Equation (3.43) follows from the assumption (2.5) in Theorem 2. Therefore, we have (3.40) converges to 0 and so does the error probability. □

Lemma 4. *Let* $(G_1, G_2) \sim \mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ *and* $R = \sum_{e\in\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1(e)=1, G_2(e)=1\}$. *If* $\boldsymbol{p}$ *satisfies constraints* (2.2) (2.3) (2.4), *and* $r = O(n\log n)$, *then*

$$\mathsf{P}(\exists\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1,G_2) \leq 0 \mid R=r) \leq 3n^2 z_6^2,$$

where $z_6 = \exp\{-\frac{2r}{n} + \frac{m}{2}\log(1-2\psi_{\mathrm{a}}) + O(1)\}$.

Proof. We will establish the above upper bound in three steps. We denote the set of vertex pairs that are *moving* under permutation $\pi^{\mathcal{E}}$ as $\mathcal{E}_{\mathrm{m}} = \{e \in \mathcal{E} : \pi^{\mathcal{E}}(e) \neq e\}$. Let

$$\tilde{R} = \sum_{e \in \mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1(e) = 1, G_2(e) = 1\}$$

represent the number of co-occurred user-user edges in $\mathcal{E}_{\mathrm{m}}$ of $G_1 \wedge G_2$. In Step 1, we apply the method of generating functions to get an upper bound on $\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = \tilde{r})$. The reason for conditioning on $\tilde{R}$ first is that the corresponding generating function only involves cycles of length $l \geq 2$ and its upper bound is easier to derive compared with the probability conditioned on R. In Step 2, we upper bound $\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid R = r)$ using result from Step 1 and properties of the Hypergeometric distribution. In Step 3, we upper bound $\mathsf{P}(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 \mid R = r)$ using the truncated union bound.

Step 1. We prove that for any $\pi \in \mathcal{S}_{n,\tilde{n}}$, $\tilde{r} = O(\frac{\tilde{t^{\mathrm{u}}} \log n}{n})$, and $z_3 = (1 - 2\psi_{\mathrm{a}})^{\frac{m}{2}}$,

$$\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = \tilde{r}\right) \leq z_3^{\tilde{n}} z_4^{\tilde{r}} z_5^{\tilde{n}} \tag{3.44}$$

for some $z_4 = O(\frac{1}{\log n})$ and some $z_5 = O(1)$.

For the induced subgraph pair on $\mathcal{E}_{\mathrm{m}} \times \mathcal{E}_{\mathrm{m}}$, define the generating function as

$$\tilde{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{y}, z) = \sum_{g \in \{0,1\}^{\mathcal{E}_{\mathrm{m}}}} \sum_{h \in \{0,1\}^{\mathcal{E}_{\mathrm{m}}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}. \tag{3.45}$$

Recall for $g, h \in \{0,1\}^{\mathcal{E}_{\mathrm{m}}}$, the expression for the extended $\delta_\pi(g,h)$, $\boldsymbol{\mu}(g,h)$ and $\boldsymbol{\nu}(g,h)$ in (3.15), (3.16) and (3.17). We have

$$\begin{aligned}\delta_\pi(g,h) = w_1 \sum_{e \in \mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{u}}} \big(\mathbb{1}\{g(e) \neq h(\pi^{\mathcal{E}}(e))\} - \mathbb{1}\{g(e) \neq h(e)\}\big) \\ + w_2 \sum_{e \in \mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{a}}} \big(\mathbb{1}\{g(e) \neq h(\pi^{\mathcal{E}}(e))\} - \mathbb{1}\{g(e) \neq h(e)\}\big).\end{aligned}$$

For the 2×2 matrices $\boldsymbol{\mu}(g,h)$ and $\boldsymbol{\nu}(g,h)$, their entries $\mu_{i,j}$ and $\nu_{i,j}$ are

$$\begin{aligned}\mu_{ij} = \mu_{ij}(g,h) = \sum_{e \in \mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{u}}} \mathbb{1}\{g(e) = i, h(e) = j\}, \\ \nu_{ij} = \nu_{ij}(g,h) = \sum_{e \in \mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{a}}} \mathbb{1}\{g(e) = i, h(e) = j\}.\end{aligned}$$

Moreover, according to the decomposition of generating function in Fact 1 and using the fact that $\mathcal{E}_{\mathrm{m}}$ only contains orbits of size larger than 1, we obtain

$$\tilde{\mathcal{A}}(\boldsymbol{x},\boldsymbol{y},z)=\prod_{l\geq 2}\mathcal{A}_l(\boldsymbol{x},z)^{t_l^{\mathrm{u}}}\prod_{l\geq 2}\mathcal{A}_l(\boldsymbol{y},z)^{t_l^{\mathrm{a}}}.$$

where t_l^{u} is the number of user-user orbits of size l and t_l^{a} is the number of user-attribute orbits of size l.

Now, by setting

$$\boldsymbol{x}=\boldsymbol{x_{11}}\triangleq\begin{pmatrix}p_{00} & p_{01}\\ p_{10} & x_{11}p_{11}\end{pmatrix}$$

and $\boldsymbol{y}=\boldsymbol{q}$, the generating function $\tilde{\mathcal{A}}(\boldsymbol{x_{11}},\boldsymbol{q},z)$ contains only two formal variables x_{11} and z. Recall the expression of $\tilde{\mathcal{A}}$ in (3.45). For each $g,h\in\{0,1\}^{\mathcal{E}_{\mathrm{m}}}$, the term in the summation of $\tilde{\mathcal{A}}(\boldsymbol{x_{11}},\boldsymbol{q},z)$ can be written as

$$\begin{aligned}&z^{\delta_\pi(g,h)}\boldsymbol{x_{11}}^{\boldsymbol{\mu}(g,h)}\boldsymbol{q}^{\boldsymbol{\nu}(g,h)}\\&=z^{\delta_\pi(g,h)}x_{11}^{\mu_{11}(g,h)}\boldsymbol{p}^{\boldsymbol{\mu}(g,h)}\boldsymbol{q}^{\boldsymbol{\nu}(g,h)}\\&=\mathsf{P}((G_1^{\mathcal{E}_{\mathrm{m}}},G_2^{\mathcal{E}_{\mathrm{m}}})=(g,h))\ \ z^{\delta_\pi(g,h)}x_{11}^{\mu_{11}(g,h)},\end{aligned}$$

where we use $G_1^{\mathcal{E}_{\mathrm{m}}}$ to denote the component of G_1 that only concerns the vertex pair set $\mathcal{E}_{\mathrm{m}}$ and thus the support of $G_1^{\mathcal{E}_{\mathrm{m}}}$ is $\{0,1\}^{\mathcal{E}_{\mathrm{m}}}$. The event $\{(G_1^{\mathcal{E}_{\mathrm{m}}},G_2^{\mathcal{E}_{\mathrm{m}}})=(g,h)\}$ is a collection of attributed graph pairs (g_1,g_2) each of which have exactly the same edges in the vertex pair set $\mathcal{E}_{\mathrm{m}}$ as (g,h).

Notice that the fixed vertex pairs $\mathcal{E}\setminus\mathcal{E}^{\mathrm{m}}$ do not have a influence on $\delta(G_1,G_2)$. The event $\{\tilde{R}=\tilde{r},\delta_\pi(G_1,G_2)=d\}$ is a collection of attributed graph pairs (g_1,g_2) such that $\mu_{11}(g_1^{\mathcal{E}_{\mathrm{m}}},g_2^{\mathcal{E}_{\mathrm{m}}})=\tilde{r}$ and $\delta_\pi(g_1^{\mathcal{E}_{\mathrm{m}}},g_2^{\mathcal{E}_{\mathrm{m}}})=d$. Then by summing over all possible $g,h\in\{0,1\}^{\mathcal{E}_{\mathrm{m}}}$, we have

$$\mathsf{P}(\delta_\pi(G_1,G_2)=d,\tilde{R}=\tilde{r})=[z^dx_{11}^{\tilde{r}}]\tilde{\mathcal{A}}(\boldsymbol{x_{11}},\boldsymbol{y},z).$$

Thus, we can write

$$\begin{aligned}&\mathsf{P}\left(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}=\tilde{r}\right)\\&=\sum_{d\leq 0}[z^dx_{11}^{\tilde{r}}]\tilde{\mathcal{A}}(\boldsymbol{x_{11}},\boldsymbol{q},z)\\&=\sum_{d\leq 0}[z^dx_{11}^{\tilde{r}}]\prod_{l\geq 2}\mathcal{A}_l(\boldsymbol{x_{11}},z)^{t_l^{\mathrm{u}}}\mathcal{A}_l(\boldsymbol{q},z)^{t_l^{\mathrm{a}}}\end{aligned}$$

$$\leq (x_{11})^{-\tilde{r}} \sum_{d\leq 0} [z^d] \prod_{l\geq 2} \mathcal{A}_l(\boldsymbol{x_{11}}, z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{q}, z)^{t_l^{\mathrm{a}}} \tag{3.46}$$

$$\leq (x_{11})^{-\tilde{r}} \prod_{l\geq 2} \mathcal{A}_l(\boldsymbol{x_{11}}, z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{q}, z)^{t_l^{\mathrm{a}}} \tag{3.47}$$

$$\leq (x_{11})^{-\tilde{r}} \mathcal{A}_2(\boldsymbol{x_{11}}, z)^{\frac{\tilde{t}^{\mathrm{u}}}{2}} \mathcal{A}_2(\boldsymbol{q}, z)^{\frac{m\tilde{n}}{2}}. \tag{3.48}$$

In (3.46), we set $x_{11} > 0$ and the inequality follows from (3.26) in Fact 3. In (3.47), we set $z \in (0,1)$ and this inequality follows from (3.27) Fact 3. Inequality in (3.48) follows from Fact 2, where

$$\tilde{t}^{\mathrm{u}} = \sum_{l\geq 2} t_l^{\mathrm{u}} l = |\mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{u}}|$$

is the number of moving user-user pairs and

$$\tilde{t}^{\mathrm{a}} = \sum_{l\geq 2} t_l^{\mathrm{a}} l = |\mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{a}}| = \tilde{n}m$$

is the number of moving user-attribute pairs.

Next, let us lower bound $\mathsf{P}(\tilde{R} = \tilde{r})$. Note that $\tilde{R} \sim \mathrm{Bin}(\tilde{t}^{\mathrm{u}}, p_{11})$. We have

$$\mathsf{P}(\tilde{R} = \tilde{r}) = \binom{\tilde{t}^{\mathrm{u}}}{\tilde{r}} p_{11}^{\tilde{r}} (1-p_{11})^{\tilde{t}^{\mathrm{u}}-\tilde{r}}$$

$$\geq \left(\frac{\tilde{t}^{\mathrm{u}} p_{11}}{\tilde{r}(1-p_{11})}\right)^{\tilde{r}} (1-p_{11})^{\tilde{t}^{\mathrm{u}}}, \tag{3.49}$$

where equation (3.49) follows since $\binom{n}{k} \geq (n/k)^k$ for any nonnegative integers $k \leq n$.

Now we combine the bounds the term in (3.48) and (3.49) to upper bound $\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = \tilde{r}\right)$. Define $p'_{ij} \triangleq \frac{p_{ij}}{1-p_{11}}$ for $i, j \in \{0, 1\}$. We have

$$\mathsf{P}\left(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = \tilde{r}\right) = \frac{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0, \tilde{R} = \tilde{r})}{\mathsf{P}(\tilde{R} = \tilde{r})}$$

$$\leq \mathcal{A}_2(\boldsymbol{q}, z)^{\frac{m\tilde{n}}{2}} \left(\frac{\tilde{r}}{x_{11} p'_{11} \tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}} \left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}}, z)}{(1-p_{11})^2}\right)^{\tilde{t}^{\mathrm{u}}/2}. \tag{3.50}$$

For the first term, similar to what we did in (3.33), we set $z = e^{-1/4}$, which satisfies the condition $z \in (0,1)$ in Fact 3. Recall the expression of $\mathcal{A}_2(\boldsymbol{y}, z)$ in (3.24), we have

$$\mathcal{A}_2(\boldsymbol{q}, z)^{\frac{m\tilde{n}}{2}}$$

$$= \left(1 + 2q_{00}q_{11}(z^{2w_2} - 1) + 2q_{10}q_{01}(z^{-2w_2} + 1)\right)^{\frac{m\tilde{n}}{2}} \tag{3.51}$$

$$= (1 - 2q_{00}q_{11} - 2q_{10}q_{01} + 4\sqrt{q_{00}q_{11}q_{10}q_{01}})^{\frac{m\tilde{n}}{2}} \tag{3.52}$$

$$= \left(1 - 2(\sqrt{q_{11}q_{00}} - \sqrt{q_{10}q_{01}})^2\right)^{\frac{m\tilde{n}}{2}}$$

$$= (1 - 2\psi_{\mathrm{a}})^{\frac{m\tilde{n}}{2}}$$

$$\triangleq z_3^{\tilde{n}}. \tag{3.53}$$

where (3.51) follows since $q_{00} + q_{01} + q_{10} + q_{11} = 1$ and (3.52) follows by plugging in $z = e^{-1/4}$ and $w_2 = \log\left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)$. For the second term in (3.50), we set

$$x_{11} = \frac{\tilde{r}\log n + p_{11}\tilde{t}^{\mathrm{u}}}{p'_{11}\tilde{t}^{\mathrm{u}}}, \tag{3.54}$$

which is positive. Then, we have

$$\left(\frac{\tilde{r}}{x_{11}p'_{11}\tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}} = \left(\frac{\tilde{r}}{\tilde{r}\log n + p_{11}\tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}} \leq \left(\frac{1}{\log n}\right)^{\tilde{r}}. \tag{3.55}$$

For the third term in (3.50), using equation (3.23) with $z = e^{-1/4}$, we have

$$\frac{\mathcal{A}_2(\boldsymbol{x_{11}}, z)}{(1-p_{11})^2}$$

$$= \frac{(1 - p_{11} + x_{11}p_{11})^2}{(1-p_{11})^2} + \frac{2x_{11}p_{11}p_{00}(\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}} - 1)}{(1-p_{11})^2} + \frac{2p_{10}p_{01}(\sqrt{\frac{p_{00}p_{11}}{p_{10}p_{01}}} - 1)}{(1-p_{11})^2}$$

$$= (1 + p'_{11}x_{11})^2 - 2x_{11}p'_{11}p'_{00} - 2p'_{10}p'_{01} + 2(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}$$

$$\leq 1 + (p'_{11}x_{11})^2 + 2p'_{11}x_{11}(p'_{10} + p'_{01}) + 2(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}},$$

where the last inequality follows since $1 - p'_{00} = p'_{10} + p'_{01}$ and $-2p'_{10}p'_{01} \leq 0$. Taking logarithm of $\left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}}, z)}{(1-p_{11})^2}\right)^{\tilde{t}^{\mathrm{u}}/2}$, we get

$$\frac{\tilde{t}^{\mathrm{u}}}{2}\log\left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}}, z)}{(1-p_{11})^2}\right) \tag{3.56}$$

$$\leq \frac{1}{2}\tilde{t}^{\mathrm{u}}(p'_{11}x_{11})^2 + \tilde{t}^{\mathrm{u}}p'_{11}x_{11}(p'_{10} + p'_{01}) + \tilde{t}^{\mathrm{u}}(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}, \tag{3.57}$$

where (3.57) follows from the inequality $\log(1 + x) \leq x$. Let us now bound the three terms in (3.57).

- For the first term, we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}(p'_{11}x_{11})^2 \\
&= \tilde{t}^{\mathrm{u}}\left(\frac{\tilde{r}\log n}{\tilde{t}^{\mathrm{u}}} + p_{11}\right)^2 \\
&= \frac{\tilde{r}^2(\log n)^2}{\tilde{t}^{\mathrm{u}}} + 2\tilde{r}(\log n)p_{11} + \tilde{t}^{\mathrm{u}}p_{11}^2 \\
&= O\left(\frac{\tilde{r}(\log n)^3}{n}\right) + 2\tilde{r}(\log n)p_{11} + \tilde{t}^{\mathrm{u}}p_{11}^2 \qquad (3.58)\\
&= O\left(\frac{\tilde{r}(\log n)^3}{n} + \frac{\tilde{r}(\log n)^2}{n} + \frac{\tilde{n}(\log n)^2}{n}\right) \qquad (3.59)\\
&= o(\tilde{r}+\tilde{n}),
\end{aligned}$$

  where (3.58) follows from the assumption $\tilde{r} = O(\frac{\tilde{t}^{\mathrm{u}}\log n}{n})$ in (3.44) and (3.59) follows since $p_{11} = O(\frac{\log n}{n})$ and $\tilde{t}^{\mathrm{u}} \le \tilde{n}n$.

- For the second term in (3.57), we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}p'_{11}x_{11}(p'_{10}+p'_{01}) \\
&= (\tilde{r}\log n + \tilde{t}^{\mathrm{u}}p_{11})(p'_{10}+p'_{01}) \\
&\le \frac{\tilde{r}(p_{10}+p_{01})\log n + \tilde{n}np_{11}(p_{10}+p_{01})}{1-p_{11}} \qquad (3.60)\\
&= O(\tilde{r}+\tilde{n}), \qquad (3.61)
\end{aligned}$$

  where (3.60) follows from $\tilde{t}^{\mathrm{u}} \le \tilde{n}n$ and (3.61) follows since $p_{01}+p_{10} = O(\frac{1}{\log n})$, $p_{11} = O(\frac{\log n}{n})$, and $1-p_{11} = \Theta(1)$.

- For the third term in (3.57), we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}} \\
&= \tilde{t}^{\mathrm{u}}\left(\frac{\tilde{r}\log n + p_{11}\tilde{t}^{\mathrm{u}}}{p'_{11}\tilde{t}^{\mathrm{u}}} + 1\right)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}} \\
&= (\tilde{r}\log n + p_{11}\tilde{t}^{\mathrm{u}} + p'_{11}\tilde{t}^{\mathrm{u}})p'_{00}\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}} \\
&\le (\tilde{r}\log n + p_{11}n\tilde{n} + p'_{11}n\tilde{n})p'_{00}\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}} \qquad (3.62)\\
&= o(\tilde{r}+\tilde{n}). \qquad (3.63)
\end{aligned}$$

Here (3.62) follows since $\tilde{t}^{\mathrm{u}} \le \tilde{n}n$. (3.63) follows since $p'_{11} = O(p_{11}) = O\left(\frac{\log n}{n}\right)$, $p'_{00} = O(1)$, and $\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\frac{1}{(\log n)^3}\right)$.

In summary, the third term of (3.50) is upper bounded as

$$\left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}}, z)}{(1-p_{11})^2}\right)^{\tilde{t}^{\mathrm{u}}/2} \le \exp\{O(\tilde{r}+\tilde{n})\}. \tag{3.64}$$

Finally, combining (3.53) (3.55) (3.64), we have

$$\begin{aligned}
&\mathsf{P}\left(\delta_\pi(G_1, G_2) \le 0 \mid \tilde{R} = \tilde{r}\right) \\
&\le (1-2\psi_{\mathrm{a}})^{\frac{m\tilde{n}}{2}} \left(\frac{1}{\log n}\right)^{\tilde{r}} \exp\{O(\tilde{r}+\tilde{n})\} \\
&\le (1-2\psi_{\mathrm{a}})^{\frac{m\tilde{n}}{2}} \left(\frac{e^{O(1)}}{\log n}\right)^{\tilde{r}} \left(e^{O(1)}\right)^{\tilde{n}} \\
&= z_3^{\tilde{n}} z_4^{\tilde{r}} z_5^{\tilde{n}}
\end{aligned}$$

for some $z_4 = O(\frac{1}{\log n})$ and $z_5 = O(1)$.

Step 2. We will prove that for any $\pi \in \mathcal{S}_{n,\tilde{n}}$ and $r = O(n \log n)$,

$$\mathsf{P}(\delta_\pi(G_1, G_2) \le 0 \mid R = r) \le z_6^{\tilde{n}} \tag{3.65}$$

for some $z_6 = \exp\{-\frac{2r}{n} + \frac{m}{2}\log(1-2\psi_{\mathrm{a}}) + O(1)\}$.

In this step, we will compute $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0 | R = r)$ through $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0 | \tilde{R} = \tilde{r})$, which involves using properties of a Hypergeometric distribution.

Recall a Hypergeometric distribution, denoted as $\mathrm{Hyp}(n, N, K)$, is the probability distribution of the number of marked elements out of the n elements we draw without replacement from a set of size N with K marked elements. Let $\Phi_{\mathrm{Hyp}}(z)$ be the probability generating function for $\mathrm{Hyp}(n, N, K)$ and $\Phi_{\mathrm{Bin}}(z)$ be the probability generating function for a binomial distribution $\mathrm{Bin}(n, \frac{K}{N})$. A few useful properties of the two distributions are as follows.

- The mean of $\mathrm{Hyp}(n, N, K)$ is nK/N.
- For all $n, N, K \in \mathbb{N}$ and $z > 0$, we have $\Phi_{\mathrm{Hyp}}(z) \le \Phi_{\mathrm{Bin}}(z)$ [2].
- $\Phi_{\mathrm{Bin}}(z) = \left(1 + \frac{K}{N}(z-1)\right)^n$.

In our problem, we are interested in the random variable $\tilde{R}|R=r$. We treat the set of moving user-user vertex pairs $\mathcal{E}_{\mathrm{u}} \cap \mathcal{E}_{\mathrm{m}}$ as a group of marked elements in $\mathcal{E}_{\mathrm{u}}$. From $\mathcal{E}_{\mathrm{u}}$, we consider drawing r vertex pairs and creating co-occurred edges for each chosen vertex pair. Along this line, the random variable $\tilde{R}|R=r$, which is the number of co-occurred edges in $\mathcal{E}_{\mathrm{m}} \cap \mathcal{E}_{\mathrm{u}}$, represents the number of marked elements out of the r chosen elements and it follows a Hypergeometric distribution $\mathrm{Hyp}(r, t^{\mathrm{u}}, \tilde{t}^{\mathrm{u}})$. From this point and on, we always consider generating functions $\Phi_{\mathrm{Hyp}}(z)$ and $\Phi_{\mathrm{Bin}}(z)$ with parameters $n=r$, $N=t^{\mathrm{u}}$, $K=\tilde{t}^{\mathrm{u}}$. Moreover, from [5, Lemma IV.5], we have the following upper bound on $\Phi_{\mathrm{Hyp}}(z)$ for any $z \in (0,1)$

$$\Phi_{\mathrm{Hyp}}(z) \leq \exp\left\{\tfrac{r\tilde{n}}{n}(-2+\tfrac{e}{n-1}+2ez)\right\}. \tag{3.66}$$

Now, we are ready for proving (3.65). We first write

$$\begin{aligned}
&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0 \mid R=r)\\
&=\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0, \tilde{R}\leq \tilde{r}^* \mid R=r)\\
&+\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0, \tilde{R}> \tilde{r}^* \mid R=r).
\end{aligned} \tag{3.67}$$

Here we set $\tilde{r}^* = C\mathsf{E}[\tilde{R} \mid R=r] = C\frac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}$, where $C>0$ is some positive constant to be specified later. Note that $t^{\mathrm{u}}=\binom{n}{2}$ and $r=O(n\log n)$ from the assumption, then we have $\tilde{r}^*=O(\frac{\tilde{t}^{\mathrm{u}}\log n}{n})$.

- For the first term in (3.67), we have

$$\begin{aligned}
&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0, \tilde{R}\leq \tilde{r}^* \mid R=r)\\
&=\sum_{\tilde{r}\leq\tilde{r}^*}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r)\ \mathsf{P}(\delta_\pi(G_1,G_2)\leq 0\mid \tilde{R}=\tilde{r}) && (3.68)\\
&\leq \sum_{\tilde{r}\leq\tilde{r}^*}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r) z_3^{\tilde{n}} z_4^{\tilde{r}} z_5^{\tilde{n}} && (3.69)\\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \sum_{\tilde{r}=0}^{n}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r) z_4^{\tilde{r}}\\
&= z_3^{\tilde{n}} z_5^{\tilde{n}} \Phi_{\mathrm{Hyp}}(z_4) && (3.70)\\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{\tfrac{\tilde{n}r}{n}(-2+\tfrac{e}{n-1}+2ez_4)\right\} && (3.71)\\
&= z_3^{\tilde{n}} (e^{O(1)})^{\tilde{n}} \exp\left\{-\tfrac{2\tilde{n}r}{n}+\tfrac{e\tilde{n}r}{n(n-1)}+O(\tfrac{1}{\log n})\tfrac{\tilde{n}r}{n}\right\} && (3.72)\\
&\leq z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(-\tfrac{2r}{n}+O(1)\right)\right\} && (3.73)
\end{aligned}$$

In (3.68), we use the conditional independence of R and $\delta_\pi(G_1, G_2)$ given $\tilde{R}$, which can be proved as follows

$$\begin{aligned}
&\mathsf{P}(\delta_\pi(G_1,G_2)\le 0|\tilde{R}=\tilde{r},R=r)\\
&=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0,\tilde{R}=\tilde{r},R=r)}{\mathsf{P}(\tilde{R}=\tilde{r},R=r)}\\
&=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0,\tilde{R}=\tilde{r},R-\tilde{R}=r-\tilde{r})}{\mathsf{P}(\tilde{R}=\tilde{r},R-\tilde{R}=r-\tilde{r})}\\
&=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0,\tilde{R}=\tilde{r})\mathsf{P}(R-\tilde{R}=r-\tilde{r})}{\mathsf{P}(\tilde{R}=\tilde{r})\mathsf{P}(R-\tilde{R}=r-\tilde{r})} \qquad (3.74)\\
&=\mathsf{P}((\delta_\pi(G_1,G_2)\le 0|\tilde{R}=\tilde{r}),
\end{aligned}$$

where (3.74) follows from the fact that $\delta_\pi(G_1, G_2)$ and $\tilde{R}$ are determined by $\mathcal{E}_\mathrm{m}$ while $R - \tilde{R}$ is determined by those fixed vertex pairs. In (3.69), we have $\tilde{r} = O(\frac{\tilde{t^\mathrm{u}} \log n}{n})$ and this inequality follows from (3.44) from Step 1. Equation (3.70) follows from the definition of the probability generating function for $\mathrm{Hyp}(r, t^\mathrm{u}, \tilde{t^\mathrm{u}})$. (3.71) follows form the conclusion about probability generating function of the hypergeometric distribution in (3.66) with $z_4 \in (0, 1)$. (3.72) is true since $z_4 = O\left(\frac{1}{\log n}\right)$ and $z_5 = O(1)$. (3.73) is true since $r = O\left(n \log n\right)$.

- For the second term of (3.67), we have

$$\begin{aligned}
&\mathsf{P}(\delta_\pi(G_1,G_2)\le 0,\tilde{R}>\tilde{r}^*\mid R=r)\\
&=\sum_{\tilde{r}>\tilde{r}^*}\mathsf{P}(\delta_\pi(G_1,G_2)\le 0,\tilde{R}=\tilde{r}\mid R=r)\\
&=\sum_{\tilde{r}>\tilde{r}^*}\mathsf{P}(\delta_\pi(G_1,G_2)\le 0\mid \tilde{R}=\tilde{r})\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r) \qquad (3.75)\\
&\le\max_{0\le\tilde{r}\le n}\{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0|\tilde{R}=\tilde{r})\}\mathsf{P}(\tilde{R}>\tilde{r}^*|R=r).
\end{aligned}$$

Here (3.75) follows from the conditional independence of $\delta_\pi(G_1, G_2)$ and R given $\tilde{R}$. To find this maximum probability, we consider the extreme case. Recall that $\delta_\pi = w_1(\Delta^\mathrm{u}(G_1, \pi(G_2)) - \Delta^\mathrm{u}(G_1, G_2)) + w_2(\Delta^\mathrm{a}(G_1, \pi(G_2)) - \Delta^\mathrm{a}(G_1, G_2))$. We have that $w_2(\Delta^\mathrm{a}(G_1, \pi(G_2)) - \Delta^\mathrm{a}(G_1, G_2))$ is independent of $\tilde{R}$. From the upper bound on generating function in (3.48), we consider $\pi^\mathcal{E}$ consisting of only 2-cycles. Since $\Delta^\mathrm{u}(G_1, \pi(G_2)) - \Delta^\mathrm{u}(G_1, G_2) > 0$ only if there exist user-user vertex

pairs such that $(G_1(e), G_2(e)) = (1,1)$ and $(G_1(\pi^{\mathcal{E}}(e))G_2(\pi^{\mathcal{E}}(e))) = (0,0)$, we have $\Delta^{\mathrm{u}}(G_1, \pi(G_2)) - \Delta^{\mathrm{u}}(G_1, G_2) \leq 0$ with probability 1 given $\tilde{R} = 0$. Therefore, given $\tilde{R} = 0$ the probability that $\delta_\pi \leq 0$ is maximized. We have

$$\begin{aligned}
&\max_{0\leq\tilde{r}\leq n} \{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = \tilde{r})\} \\
&\leq \mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid \tilde{R} = 0) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}}, \qquad (3.76)
\end{aligned}$$

where (3.76) follows from (3.44) in Step 1 with $\tilde{r} = 0$. Now we get

$$\begin{aligned}
&\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0, \tilde{R} > \tilde{r}^* \mid R = r) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \mathsf{P}(\tilde{R} > \tilde{r}^* \mid R = r) \\
&= z_3^{\tilde{n}} z_5^{\tilde{n}} \sum_{i>\tilde{r}^*} [z^i] \Phi_{\mathrm{Hyp}}(z) \qquad (3.77) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*} \Phi_{\mathrm{Hyp}}(z) \qquad (3.78) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*} \Phi_{\mathrm{Bin}}(z) \qquad (3.79) \\
&= z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*} \left(1 + \tfrac{\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(z-1)\right)^r \qquad (3.80) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*} \exp\left\{\tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(z-1)\right\} \qquad (3.81) \\
&= z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{-\tilde{r}^* + \tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(e-1)\right\} \qquad (3.82) \\
&= z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{\tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(-C-1+e)\right\} \qquad (3.83) \\
&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{\tfrac{r\tilde{n}(n-2)}{n(n-1)}(-C-1+e)\right\} \qquad (3.84) \\
&\leq z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(\tfrac{r}{n}(-C-1+e) + O(1)\right)\right\} \qquad (3.85) \\
&= o\left(z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(-\tfrac{2r}{n} + O(1)\right)\right\}\right) \qquad (3.86)
\end{aligned}$$

In (3.77), $\Phi_{\mathrm{Hyp}}(z)$ is a probability generating function for $\mathrm{Hyp}(r, t^{\mathrm{u}}, \tilde{t}^{\mathrm{u}})$. In (3.78), we set $z > 1$ and the inequality follows from (3.28) in Fact 3. In (3.79), $\Phi_{\mathrm{Bin}}(z)$ is a probability generating function for $\mathrm{Bin}(r, \frac{\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}})$ and this inequality follows from the property of a Hypergeometric distribution. (3.80) follows from the definition of $\Phi_{\mathrm{Bin}}(z)$. (3.81) follows form the inequality $1 + x \leq e^x$. In (3.82), we set $z = e$. In (3.83), we plug in $\tilde{r}^* = C\frac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}$ where C is larger than $(e-1)$. In (3.84), we use the relation $\tilde{t}^{\mathrm{u}} \geq \frac{\tilde{n}(n-2)}{2}$ from (3.35) and $t^{\mathrm{u}} = \binom{n}{2}$. In (3.85), we plug in $z_5 = O(1)$. (3.86) is true because we can always find $C > e + 1$ such that (3.85) is exponentially smaller than (3.73).

We conclude that the second term of (3.67) is negligible compared with the upper bound of the first term given in (3.73). Combining the two terms, (3.67) can be bounded as

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\le 0 \mid R=r)\\ &\le \exp\left\{\tilde{n}\left(-\tfrac{2r}{n}+\tfrac{m}{2}\log(1-2\psi_{\mathrm{a}})+O(1)\right)\right\}\\ &= z_6^{\tilde{n}}.\end{aligned}$$

Step 3. We now establish the desired error bound

$$\mathsf{P}(\exists \pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\},\delta_\pi(G_1,G_2)\le 0 \mid R=r)\le 3n^2z_6^2,$$

where $z_6=\exp\{-\frac{2r}{n}+\frac{m}{2}\log(1-2\psi_{\mathrm{a}})+O(1)\}$.

When $nz_6>2/3$, we have

$$\mathsf{P}(\exists \pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\},\delta_\pi(G_1,G_2)\le 0 \mid R=r)\le 1\le 3n^2z_6^2.$$

Now assume that $nz_6\le 2/3$. We can bound

$$\begin{aligned}&\mathsf{P}(\exists \pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\},\delta_\pi(G_1,G_2)\le 0 \mid R=r)\\ &\le \sum_{\tilde{n}=2}^{n}\sum_{\pi\in\mathcal{S}_{n,\tilde{n}}}\mathsf{P}(\delta_\pi(G_1,G_2)\le 0 \mid R=r) &(3.87)\\ &\le \sum_{\tilde{n}=2}^{n}|\mathcal{S}_{n,\tilde{n}}|\max_{\pi\in\mathcal{S}_{n,\tilde{n}}}\{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0 \mid R=r)\}\\ &\le \sum_{\tilde{n}=2}^{n}|\mathcal{S}_{n,\tilde{n}}|z_6^{\tilde{n}} &(3.88)\\ &\le \sum_{\tilde{n}=2}^{n}n^{\tilde{n}}z_6^{\tilde{n}} &(3.89)\\ &\le \frac{(nz_6)^2}{1-nz_6}\\ &\le 3n^2z_6^2, &(3.90)\end{aligned}$$

where (3.87) follows from the union bound, (3.88) follows from inequality (3.65) proved in Step 2, (3.89) follows since $|\mathcal{S}_{n,\tilde{n}}|\le n^{\tilde{n}}$, and (3.90) holds since $nz_6\le 2/3$.

In summary, $3n^2z_6^2$ is always an upper bound on the conditional probability. This completes the proof of Lemma 4. □

Chapter 4

Proof of the Converse

In this chapter, we give a detailed proof for Theorem 3. Let (G_1, G_2) be an attributed graph pair generated from $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. In this proof, we focus on the intersection graph of G_1 and G_2, which is denoted as $G_1 \wedge G_2$. The intersection graph has vertex set $\mathcal{V} = \mathcal{V}_{\mathrm{u}} \cup \mathcal{V}_{\mathrm{a}}$ and its edge set is the intersection of the edge sets of G_1 and G_2. We say a permutation π on the vertex set $\mathcal{V}$ is an automorphism of $G_1 \wedge G_2$ if π is *edge-preserving*, i.e., a vertex pair (i, j) is in the edge set of $G_1 \wedge G_2$ if and only if $(\pi(i), \pi(j))$ is in the edge set of $G_1 \wedge G_2$. Note that the identity permutation is always an automorphism. We use $\mathrm{Aut}(G_1 \wedge G_2)$ to denote the set of automorphisms of $G_1 \wedge G_2$. By Lemma 5 below, we can further argue that exact alignment cannot be achieved w.h.p. if $\mathrm{Aut}(G_1 \wedge G_2)$ contains permutations other than the identity permutation. Along this line, we establish the condition for not achieving exact alignment w.h.p. by analyzing automorphisms of $G_1 \wedge G_2$.

Lemma 5 ([4]). *Let (G_1, G_2) be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. Given $|\mathrm{Aut}(G_1 \wedge G_2)|$, the probability that MAP estimator succeeds is at most $\frac{1}{|\mathrm{Aut}(G_1 \wedge G_2)|}$.*

In the proof of Theorem 3, we focus on the automorphisms given by swapping two user vertices. To this end, we first define the following equivalence relation between a pair of user vertices. We say two user vertices i and j $(i \neq j)$ are *indistinguishable* in $G_1 \wedge G_2$, denoted as $i \equiv j$, if $(G_1 \wedge G_2)((i, v)) = (G_1 \wedge G_2)((j, v))$ for all $v \in \mathcal{V}$. It is not hard to see that swapping two indistinguishable vertices is an automorphism of $G_1 \wedge G_2$, and thus $|\mathrm{Aut}(G_1 \wedge G_2) \backslash \{\text{identity permutation}\}| \geq |\{\text{indistinguishable vertex pairs}\}|$. Therefore, in the proof below, we show that the number of such indistinguishable vertex pairs is positive with a large probability, which further implies that $|\mathrm{Aut}(G_1 \wedge G_2)| \geq 2$ with a large probability and eventually proves Theorem 3.

Theorem 3 (Converse). *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}, m, \boldsymbol{q})$. If*

$$np_{11} + mq_{11} - \log n \to -\infty, \tag{2.6}$$

then no algorithm guarantees exact alignment w.h.p.

Proof of Theorem 3. Let G_1 and G_2 be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ and let $G = G_1 \wedge G_2$. Let X denote the number of indistinguishable user vertex pairs in G, i.e.,

$$X = \sum_{i<j\colon i,j\in\mathcal{V}_\mathrm{u}} \mathbb{1}\{i \equiv j\}.$$

We will show that $\mathsf{P}(X = 0) \to 0$ as $n \to \infty$ if the condition (2.6) in Theorem 3 is satisfied.

We start by upper-bounding $\mathsf{P}(X = 0)$ using Chebyshev's inequality

$$\mathsf{P}(X = 0) \le \frac{\mathsf{Var}(X)}{\mathsf{E}[X]^2} = \frac{\mathsf{E}[X^2] - \mathsf{E}[X]^2}{\mathsf{E}[X]^2}. \tag{4.1}$$

For the first moment term $\mathsf{E}[X]$, we have

$$\mathsf{E}[X] = \sum_{i<j} \mathsf{P}(i \equiv j) = \binom{n}{2}\mathsf{P}(i \equiv j). \tag{4.2}$$

For the second moment term $\mathsf{E}[X^2]$, we expand the sum as

$$\begin{aligned}
\mathsf{E}[X^2] &= \mathsf{E}\left[\sum_{i<j} \mathbb{1}\{i \equiv j\} \cdot \sum_{k<l} \mathbb{1}\{k \equiv l\}\right] \\
&= \mathsf{E}\Bigg[\sum_{i<j} \mathbb{1}\{i \equiv j\} + \sum_{\substack{i,j,k,l\colon i<j,k<l \\ i,j,k,l \text{ are distinct}}} \mathbb{1}\{i \equiv j\} \cdot \mathbb{1}\{k \equiv l\} \\
&\qquad + \sum_{\substack{i,j,k,l\colon i<j,k<l \\ \{i,j\} \text{ and } \{k,l\} \text{ share one element}}} \mathbb{1}\{i \equiv j \equiv k \equiv l\}\Bigg] \\
&= \binom{n}{2}\mathsf{P}(i \equiv j) + \binom{n}{4}\binom{4}{2}\mathsf{P}(i \equiv j \text{ and } k \equiv l) + 6\binom{n}{3}\mathsf{P}(i \equiv j \equiv k),
\end{aligned} \tag{4.3}$$

where i, j, k, l are distinct in (4.3). With (4.2) and (4.3), the upper bound given by Chebyshev's inequality in (4.1) can be written as

$$\frac{\mathsf{Var}(X)}{\mathsf{E}[X]^2} = \frac{2}{n(n-1)\mathsf{P}(i \equiv j)} + \frac{4(n-2)}{n(n-1)}\frac{\mathsf{P}(i \equiv j \equiv k)}{\mathsf{P}(i \equiv j)^2}$$

$$+ \frac{(n-2)(n-3)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2} - 1. \tag{4.4}$$

To compute $\mathsf{P}(i \equiv j)$, we look into the event $\{i \equiv j\}$ which is the intersection of A_1 and A_2, where $A_1 = \{\forall v \in \mathcal{V}_{\mathrm{u}} \setminus \{i,j\}, G((i,v)) = G((j,v))\}$, and $A_2 = \{\forall u \in \mathcal{V}_{\mathrm{a}}, G((i,u)) = G((j,u))\}$. Recall that in the intersection graph $G = G_1 \wedge G_2$, the edge probability is p_{11} for user-user pairs and q_{11} for user-attribute pairs. Therefore,

$$\begin{aligned}
\mathsf{P}(A_1) &= \sum_{i=0}^{n-2} \binom{n-2}{i} p_{11}^{2i} (1-p_{11})^{2(n-2-i)} \\
&= \left(p_{11}^2 + (1-p_{11})^2\right)^{n-2}, \\
\mathsf{P}(A_2) &= \sum_{i=0}^{m} \binom{m}{i} p_{11}^{2i} (1-p_{11})^{2(m-i)} \\
&= \left(q_{11}^2 + (1-q_{11})^2\right)^{m}.
\end{aligned}$$

Since A_1 and A_2 are independent, we have

$$\begin{aligned}
&\mathsf{P}(i \equiv j) \qquad (4.5) \\
&= \mathsf{P}(A_1)\mathsf{P}(A_2) \\
&= \left(p_{11}^2 + (1-p_{11})^2\right)^{n-2} \left(q_{11}^2 + (1-q_{11})^2\right)^{m} \\
&= \left(1 - 2p_{11} + 2p_{11}^2\right)^{n-2} \left(1 - 2q_{11} + 2q_{11}^2\right)^{m}. \qquad (4.6)
\end{aligned}$$

Similarly, to compute $\mathsf{P}(i \equiv j \equiv k)$, we look into the event $\{i \equiv j \equiv k\}$ which is the intersection of events B_0, B_1 and B_2, where $B_0 = \{G((i,j)) = G((j,k)) = G((i,k))\}$, $B_1 = \{\forall v \in \mathcal{V}_{\mathrm{u}} \setminus \{i,j,k\}, G((i,v)) = G((j,v)) = G((k,v))\}$, and $B_2 = \{\forall u \in \mathcal{V}_{\mathrm{a}}, G((i,u)) = G((j,u)) = G((k,u))\}$. Then, the probabilities of those three events are

$$\begin{aligned}
\mathsf{P}(B_0) &= p_{11}^3 + (1-p_{11})^3, \\
\mathsf{P}(B_1) &= \left(p_{11}^3 + (1-p_{11})^3\right)^{n-3}, \\
\mathsf{P}(B_2) &= \left(q_{11}^3 + (1-q_{11})^3\right)^{m}.
\end{aligned}$$

Since the events B_0, B_1 and B_2 are independent, we have

$$\begin{aligned}
&\mathsf{P}(i \equiv j \equiv k) \\
&= \mathsf{P}(B_0)\mathsf{P}(B_1)\mathsf{P}(B_2) \\
&= \left(p_{11}^3 + (1-p_{11})^3\right)^{n-2} \left(q_{11}^3 + (1-q_{11})^3\right)^{m}
\end{aligned}$$

$$= (1 - 3p_{11} + 3p_{11}^2)^{n-2}(1 - 3q_{11} + 3q_{11}^2)^m.$$

To compute $\mathsf{P}(i \equiv j \text{ and } k \equiv l)$, we look into the event $\{i \equiv j \text{ and } k \equiv l\}$ which is the intersection of C_0, C_1, C_1', C_2 and C_2', where $C_0 = \{G(i,k) = G(j,k) = G(i,l) = G(j,l)\}$, $C_1 = \{\forall v \in \mathcal{V}_{\mathrm{u}} \setminus \{i,j,k,l\}, G(i,v) = G(j,v)\}$, $C_1' = \{\forall v \in \mathcal{V}_{\mathrm{u}} \setminus \{i,j,k,l\}, G(k,v) = G(l,v)\}$, $C_2 = \{\forall u \in \mathcal{V}_{\mathrm{a}}, G(i,u) = G(j,u)\}$ and $C_2' = \{\forall u \in \mathcal{V}_{\mathrm{a}}, G(k,u) = G(l,u)\}$. The probabilities of those events are

$$\begin{aligned}
\mathsf{P}(C_0) &= p_{11}^6 + p_{11}^4(1-p_{11})^2 + p_{11}^2(1-p_{11})^4 + (1-p_{11})^6, \\
\mathsf{P}(C_1) &= \mathsf{P}(C_1') = (p_{11}^2 + (1-p_{11})^2)^{n-4}, \\
\mathsf{P}(C_2) &= \mathsf{P}(C_2') = (q_{11}^2 + (1-q_{11})^2)^m.
\end{aligned}$$

Since C_0, C_1, C_1', C_2 and C_2' are independent, we have

$$\begin{aligned}
&\mathsf{P}(i \equiv j \text{ and } k \equiv l) \\
&= \mathsf{P}(C_0)\mathsf{P}(C_1)\mathsf{P}(C_1')\mathsf{P}(C_2)\mathsf{P}(C_2') \\
&= \mathsf{P}(C_0)(p_{11}^2 + (1-p_{11})^2)^{2n-8}(q_{11}^2 + (1-q_{11})^2)^{2m}.
\end{aligned}$$

Now we are ready to analyze the terms in (4.4). For the last two terms, note that $\frac{(n-2)(n-3)}{n(n-1)} \to 1$ and $\frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2} \to 1$ because $p_{11} < \frac{\log n}{n}$ from the condition (2.6). Therefore, we have $\frac{(n-2)(n-3)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2} - 1 \to 0$ as $n \to \infty$. Then we just need to bound the first two terms in (4.4). For the first term $\frac{2}{n(n-1)\mathsf{P}(i \equiv j)}$, plugging in the expression in (4.6) gives

$$\begin{aligned}
&-\log \frac{2}{n(n-1)\mathsf{P}(i \equiv j)} \\
&= 2\log n + (n-2)\log(1 - 2p_{11} + 2p_{11}^2) + m\log(1 - 2q_{11} + 2q_{11}^2) + O(1) \\
&\geq 2\log n - 2np_{11} - 2mq_{11} + O(1) \qquad (4.7) \\
&= \omega(1). \qquad (4.8)
\end{aligned}$$

Here (4.7) follows from the inequality $\log(1 - 2x + 2x^2) \geq -2x$ for any $x \in [0,1]$, which can be verified by showing that function $f_1(x) = \log(1 - 2x + 2x^2) + 2x$ is monotone increasing in [0,1] and thus $f_1(x) \geq f_1(0) = 0$. Equation (4.8) follows from the condition (2.6) in Theorem 3. Therefore, the first term in (4.4) $\frac{2}{n(n-1)\mathsf{P}(i \equiv j)} \to 0$ as $n \to \infty$.

Next, for the second term $\frac{4(n-2)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \equiv k)}{\mathsf{P}(i \equiv j)^2}$ in (4.4), we have

$$-\log\left(\frac{4(n-2)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \equiv k)}{\mathsf{P}(i \equiv j)^2}\right)$$

$$= \log n - (n-2)\log\left(\frac{1-3p_{11}+3p_{11}^2}{(1-2p_{11}+2p_{11}^2)^2}\right) - m\log\left(\frac{1-3q_{11}+3q_{11}^2}{(1-2q_{11}+2q_{11}^2)^2}\right) + O(1)$$

$$\geq \log n - np_{11} - mq_{11} + O(1) \tag{4.9}$$

$$= \omega(1). \tag{4.10}$$

Here (4.9) follows from the inequality $\log\left(\frac{1-3x+3x^2}{(1-2x+2x^2)^2}\right) \leq x$ for any $x \in [0,1]$, which can be verified by showing that the function $f_2(x) = \log\left(\frac{1-3x+3x^2}{(1-2x+2x^2)^2}\right) - x$ is monotone decreasing in $[0,1]$ and thus $f_2(x) \leq f_2(0) = 0$. Equation (4.10) follows from the condition (2.6) in Theorem 3. Hence, the second term in (4.4) also converges to 0 as $n \to \infty$, which completes the proof for $\mathsf{P}(X=0) \to 0$ as $n \to \infty$.

Now we derive an upper bound on the probability of exact alignment under the MAP estimator, which is also an upper bound for any estimator since MAP minimizes the probability of error. Note that by Lemma 5, $\mathsf{P}(\pi_{\text{MAP}} = \Pi^* | X = x) \leq \frac{1}{x+1}$, which is at most $1/2$ when $x \geq 1$. Therefore,

$$\begin{aligned}
\mathsf{P}(\pi_{\text{MAP}} = \Pi^*) &= \mathsf{P}(\pi_{\text{MAP}} = \Pi^* | X = 0)\mathsf{P}(X=0) \\
&\quad + \mathsf{P}(\pi_{\text{MAP}} = \Pi^* | X \geq 1)\mathsf{P}(X \geq 1) \\
&\leq \mathsf{P}(X=0) + \frac{1}{2}\mathsf{P}(X \geq 1) \\
&= \frac{1}{2} + \frac{1}{2}\mathsf{P}(X=0),
\end{aligned}$$

which goes to $1/2$ as $n \to \infty$ and thus is bounded away from 1. This completes the proof that no algorithm can guarantee exact alignment w.h.p. □

Chapter 5

Concluding Remarks

In this thesis, we focus on studying the attributed graph alignment problem. We contribute mainly from three perspectives: propose the attributed Erdős–Rényi pair model, characterize the information-theoretic limits on exact alignment, and specialize our results for understanding three other well-studied graph alignment models. The current limitation is that there is still a gap between our achievability results and the converse result. In this chapter, we highlight several extensions and potential future directions for our work.

A potential improvement on the converse: In our proof of the converse result (see Chapter 4), we examine the existence of indistinguishable vertex pairs, which leads to a failure of the MAP estimator. In a very recent study [27], the authors consider a different error event, which includes our error event about the indistinguishable vertex pair as a special case. From their more general error event, the authors prove the converse of Erdős–Rényi graph alignment problem. Inspired by this, we tried an easy extension of their technique and generalized our error event about the indistinguishable vertex pair under the attributed Erdős–Rényi setting. As a result, we are able to demonstrate that when $np_{11}p_{00} + mq_{11}q_{00} \leq \log n - \omega(1)$, the MAP estimator cannot achieve exact alignment with high probability. We defer the detailed proof to Appendix E. This new result extend our converse region in Theorem 3, and also imply the possibility of further proving an enlarged converse region by implementing the entire idea of [27].

The interplay between graph alignment and graph clustering: Alignment and clustering of graphs are both important topics in graph structure data science research. While graph alignment aims at recovering the vertex correspondence of multiple graphs, graph clustering is concerned with recovering the community structure of a single graph. There has recently been a growing interest in combining the two methods. There are two distinct ways to combine them: using graph alignment to assist graph clustering, or using the graph clustering method to assist graph alignment. For example, in [12], the authors investigate the problem of community recovery from two graphs generated from the correlated Stochastic Block Model [16].

They present an approach that first partially aligns two graphs, and then performs community recovery for both within and outside the aligned vertex set. The feasible regions of their algorithm include an interesting case where exact community recovery is not achievable using a single graph but achievable using a pair of correlated graphs.

Our study on the attributed graph alignment sheds light on the second idea of combining—using the graph clustering methods to assist graph alignment. Let us consider such a setting: a graph pair is generated from the correlated Stochastic Block Model, where the community labels of the two graphs ares recovered already using graph cluster method as a pre-processing step. The community labels provide a natural partition in the graphs, which, combining with our study of attributed graph model, eventually allows us to perform graph alignment in a step-by-step manner. For example, we may consider the following algorithm: out of the several communities, we first perform graph alignment only on the most strongly correlated community. Then we treat vertices in the the aligned community as attributes and perform attributed alignment on the second strongly correlated community. This procedure can be repeated iteratively for multiple community graphs. The intuition behind this is that knowing the community structure enables us to determine an "easy to hard" order for aligning those subgraphs. More specifically, strongly correlated subgraphs are easier to align, so we align them first. The strategy of treating the aligned vertices as attributes allows us to incorporate side information on less correlated subgraphs, and thus make the later alignment steps easier.

Correlation between attributes and graph structure: Our attributed Erdős–Rényi model is initially motivated by the existence of side information associated with individual vertices in real-world networks, such as user profiles from social networks. In our model formulation, we assumed that the user-attribute edges are independent of the user-user edges, which is not necessarily the case in practice. In the social network example, users studying at the same university are more likely to be friends than users attending different universities. This observation suggests an improvement in the random graph models – the correlation between attribute information and the graph structure should also be captured. Under such new model formulations, it would be practically meaningful to investigate both the information-theoretic limits and design efficient algorithms for graph alignment. Although there is no existing work on graph alignment under these models, there are several recently proposed attributed graph models that capture the correlation between attributes and graph structure. For example, in [14], the authors proposed a random graph model, named the

multiplicative attribute graph model, where the probability of a user-user edge depends on the product of individual attribute-attribute similarity. Models like this could bring more practical relevance in future graph alignment research.

Bibliography

[1] M. Cho and K. M. Lee. Progressive graph matching: Making a move of graphs via probabilistic voting. In *Proc. IEEE Comput. Vision and Pattern Recognit.*, pages 398–405, 2012.

[2] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25:285–287, 1979.

[3] D. Cullina, P. Mittal, and N. Kiyavash. Fundamental limits of database alignment. In *Proc. IEEE Int. Symp. Information Theory*, pages 651–655, 2018.

[4] Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for Erdős-Rényi graph matching. *ACM SIGMETRICS Perform. Evaluation Rev.*, 44(1):63–72, 2016.

[5] Daniel Cullina and Negar Kiyavash. Exact alignment recovery for correlated Erdős-Rényi graphs. *arXiv:1711.06783 [cs.IT]*, 2017.

[6] Daniel Cullina, Prateek Mittal, and Negar Kiyavash. Fundamental limits of database alignment. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 651–655, 2018.

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

[8] Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. 2020.

[9] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations: Algorithm and theory. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2985–2995. PMLR, 13–18 July 2020.

[10] Donniell E Fishkind, Sancar Adali, Heather G Patsolic, Lingyao Meng, Digvijay Singh, Vince Lyzinski, and Carey E Priebe. Seeded graph matching. *Pattern recognition*, 87:203–215, 2019.

[11] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[12] Julia Gaudio, Miklos Z Racz, and Anirudh Sridhar. Exact community recovery in correlated stochastic block models. *arXiv preprint arXiv:2203.15736*, 2022.

[13] Aria D. Haghighi, Andrew Y. Ng, and Christopher D. Manning. Robust textual inference via graph matching. In *Human Lang. Technol. and Empirical Methods in Natural Lang. Process.*, 2005.

[14] Myunghwan Kim and Jure Leskovec. Multiplicative attribute graph model of real-world networks. *Internet mathematics*, 8(1-2):113–160, 2012.

[15] Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.*, 7(5):377–388, January 2014.

[16] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[17] Joseph Lubars and R Srikant. Correcting the output of approximate graph matching algorithms. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1745–1753. IEEE, 2018.

[18] Vince Lyzinski, Donniell E Fishkind, and Carey E Priebe. Seeded graph matching for correlated erdös-rényi graphs. *J. Mach. Learn. Res.*, 15(1):3513–3540, 2014.

[19] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Exact matching of random graphs with constant correlation. 2021.

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

[21] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proc. IEEE Symp. Security and Privacy*, pages 173–187, 2009.

[22] 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, 2008.

[23] Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 1235–1243, 2011.

[24] F. Shirani, S. Garg, and E. Erkip. A concentration of measure approach to database de-anonymization. In *Proc. IEEE Int. Symp. Information Theory*, pages 2748–2752, 2019.

[25] 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.

[26] Ziao Wang, Ning Zhang, Weina Wang, and Lele Wang. On the feasible region of efficient algorithms for attributed graph alignment. *arXiv preprint arXiv:2201.10106*, 2022.

[27] Yihong Wu, Jiaming Xu, and H Yu Sophie. Settling the sharp reconstruction thresholds of random graph matching. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2714–2719. IEEE, 2021.

[28] 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.

[29] Ning Zhang, Weina Wang, and Lele Wang. Attributed graph alignment. *arXiv:2102.00665 [cs.IT]*, 2021.

[30] Si Zhang and Hanghang Tong. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1345–1354, 2016.

[31] Si Zhang and Hanghang Tong. Attributed network alignment: Problem definitions and fast solutions. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1680–1692, 2018.

[32] Qinghai Zhou, Liangyue Li, Xintao Wu, Nan Cao, Lei Ying, and Hanghang Tong. Attent: Active attributed network alignment. In *Proceedings of the Web Conference 2021*, pages 3896–3906, 2021.

Appendix A

MAP Estimator

Lemma 2 (MAP estimator)**.** *Let (G_1, G_2') be an observable pair generated from the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. The MAP estimator of the permutation Π^* based on (G_1, G_2') simplifies to*

$$\begin{aligned}&\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') \\ &= \operatorname*{argmin}_{\pi \in \mathcal{S}_n} \{w_1 \Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + w_2 \Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))\},\end{aligned}$$

where $w_1 = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)$, $w_2 = \log\left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)$, *and*

$$\begin{aligned}\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) &= \sum_{(i,j)\in\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1((i,j)) \neq G_2'((\pi(i), \pi(j)))\}, \\ \Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2')) &= \sum_{(i,v)\in\mathcal{E}_{\mathrm{a}}} \mathbb{1}\{G_1((i,v)) \neq G_2'((\pi(i), v))\}.\end{aligned}$$

Proof. Let (g_1, g_2') be a realization of an observable pair (G_1, G_2') from $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. Then the posterior of the permutation Π^* can be written as:

$$\begin{aligned}&\mathsf{P}(\Pi^* = \pi | G_1 = g_1, G_2' = g_2') \\ &= \frac{\mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi)\mathsf{P}(\Pi^* = \pi)}{P(G_1 = g_1, G_2' = g_2')} \\ &\propto \mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi) \qquad \text{(A.1)} \\ &= \mathsf{P}(G_1 = g_1, G_2 = \pi^{-1}(g_2')) \qquad \text{(A.2)} \\ &= \prod_{(i,j)\in\{0,1\}^2} p_{ij}^{\mu_{ij}(g_1, \pi^{-1}(g_2'))} q_{ij}^{\nu_{ij}(g_1, \pi^{-1}(g_2'))}. \qquad \text{(A.3)}\end{aligned}$$

Here equation (A.1) follows from the fact that Π^* is uniformly drawn from $\mathcal{S}_n$ and $\mathsf{P}(G_1 = g_1, G_2' = g_2')$ does not depend on π. Equation (A.2) is due to the independence between Π^* and (G_1, G_2).

To further simplify equation (A.3), note that the total number of edges in a graph is invariant under any permutation. We define $\beta^{\mathrm{u}}(G_1)$ as the

total number of user-user edges in graph G_1 and $\beta^{\mathrm{u}}(\pi^{-1}(G_2'))$ for graph $\pi^{-1}(G_2')$. Similarly, we define $\beta^{\mathrm{a}}(G_1)$ and $\beta^{\mathrm{a}}(\pi^{-1}(G_2'))$ as the total number of user-attribute edges for graph G_1 and $\pi^{-1}(G_2')$, respectively. Recall our definitions on Hamming distance $\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))$ and $\boldsymbol{\mu}(G_1, \pi^{-1}(G_2'))$, and notice that $\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) = \mu_{10} + \mu_{01}$. Moreover, we have $\beta^{\mathrm{u}}(G_1) = \mu_{11} + \mu_{10}$ and $\beta^{\mathrm{u}}(G_2) = \beta^{\mathrm{u}}(\pi^{-1}(G_2')) = \mu_{11} + \mu_{01}$. Then, for the user-user set $\mathcal{E}_{\mathrm{u}}$, we have

$$\begin{aligned}
\mu_{11} &= \frac{\beta^{\mathrm{u}}(G_1) + \beta^{\mathrm{u}}(\pi^{-1}(G_2'))}{2} - \frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2} \\
\mu_{10} &= \frac{\beta^{\mathrm{u}}(G_1) - \beta^{\mathrm{u}}(\pi^{-1}(G_2'))}{2} + \frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2} \\
\mu_{01} &= \frac{\beta^{\mathrm{u}}(\pi^{-1}(G_2')) - \beta^{\mathrm{u}}(G_1)}{2} + \frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2} \\
\mu_{00} &= \binom{n}{2} - \frac{\beta^{\mathrm{u}}(G_1) + \beta^{\mathrm{u}}(\pi^{-1}(G_2'))}{2} - \frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2}.
\end{aligned}$$

Similarly, for the user-attribute set $\mathcal{E}_{\mathrm{a}}$, we have $\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2')) = \nu_{10} + \nu_{01}$, $\beta^{\mathrm{a}}(G_1) = \nu_{11} + \nu_{10}$ and $\beta^{\mathrm{a}}(G_2) = \beta^{\mathrm{a}}(\pi^{-1}(G_2')) = \nu_{11} + \nu_{01}$. Therefore, we get

$$\begin{aligned}
\nu_{11} &= \frac{\beta^{\mathrm{a}}(G_1) + \beta^{\mathrm{a}}(\pi^{-1}(G_2'))}{2} - \frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2} \\
\nu_{10} &= \frac{\beta^{\mathrm{a}}(G_1) - \beta^{\mathrm{a}}(\pi^{-1}(G_2'))}{2} + \frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2} \\
\nu_{01} &= \frac{\beta^{\mathrm{a}}(\pi^{-1}(G_2')) - \beta^{\mathrm{a}}(G_1)}{2} + \frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2} \\
\nu_{00} &= nm - \frac{\beta^{\mathrm{a}}(G_1) + \beta^{\mathrm{a}}(\pi^{-1}(G_2'))}{2} - \frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2}.
\end{aligned}$$

Since $\beta^{\mathrm{u}}(G_1), \beta^{\mathrm{u}}(\pi^{-1}(G_2')), \beta^{\mathrm{a}}(G_1)$, and $\beta^{\mathrm{a}}(\pi^{-1}(G_2'))$ do not depend on π, we can further simplify the posterior as follows

$$\begin{aligned}
&\mathsf{P}(\Pi^* = \pi | G_1 = G_1, G_2' = G_2') \\
&\propto \prod_{(i,j)\in\{0,1\}^2} p_{ij}^{\mu_{ij}(G_1, \pi^{-1}(G_2'))} q_{ij}^{\nu_{ij}(G_1, \pi^{-1}(G_2'))} \\
&\propto \left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)^{-\frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2}} \left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)^{-\frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2}} \qquad \text{(A.4)} \\
&= \exp\left\{-w_1 \frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2} - w_2 \frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2}\right\}, \qquad \text{(A.5)}
\end{aligned}$$

where $w_1 \triangleq \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)$ and $w_2 \triangleq \log\left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)$. Note that $w_1 > 0$ and $w_2 > 0$ since we assume that the edges in G_1 and G_2 are positively correlated. Therefore, of all the permutation in S_n, the one that minimizes the weighted Hamming distance $w_1\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + w_2\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))$ achieves the maximum posterior probability. □

Now consider the seeded Erdős–Rényi pair model $\mathcal{G}(n, m, \boldsymbol{p})$. Recall that when specializing the attributed Erdős–Rényi pair model by setting $\boldsymbol{p} = \boldsymbol{q}$, we can treated the m attributes as m seeds. The only difference between the $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{p})$ model and the seeded model $\mathcal{G}(n, m, \boldsymbol{p})$ is that there are no edges between seeds in the specialized model, but those edges exist in the seeded model. Here, we show that this distinction has no influence on the information-theoretic limit of exact alignment. To see this, we prove that the optimal estimator – MAP estimator still simplifies to minimizing the Hamming distance of the user-user edges and user-seed edges.

Lemma 6. *Let (G_1, G_2') be an pair seeded graphs generated from the seeded Erdős–Rényi pair $\mathcal{G}(n, m, \boldsymbol{p})$. The MAP estimator of the permutation Π^* based on (G_1, G_2') simplifies to*

$$\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') = \operatorname*{argmin}_{\pi \in \mathcal{S}_n} \{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + \Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))\},$$

where

$$\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) = \sum_{(i,j)\in\mathcal{E}_{\mathrm{u}}} \mathbb{1}\{G_1((i,j)) \neq G_2'((\pi(i), \pi(j)))\},$$
$$\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2')) = \sum_{(i,v)\in\mathcal{E}_{\mathrm{a}}} \mathbb{1}\{G_1((i,v)) \neq G_2'((\pi(i), v))\}.$$

Proof. To start, we have the posterior of the underlying permutation.

$$\begin{aligned}
&\mathsf{P}(\Pi^* = \pi | G_1 = g_1, G_2' = g_2') \\
&= \frac{\mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi)\mathsf{P}(\Pi^* = \pi)}{P(G_1 = g_1, G_2' = g_2')} \\
&\propto \mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi) \qquad \text{(A.6)} \\
&= \mathsf{P}(G_1 = g_1, G_2 = \pi^{-1}(g_2')). \qquad \text{(A.7)}
\end{aligned}$$

Here (A.6) follows since Π^* is uniformly drawn. (A.7) follows since Π^* is independent of G_1 and G_2. For ease of notation, we use g_2^π to denote $\pi^{-1}(g_2')$. Then according to the seeded graph model in Chapter 1.3, we have

$$\mathsf{P}(G_1 = g_1, G_2 = g_2^\pi)$$

$$= p_{11}^{\mu_{11}(g_1,g_2^\pi)} p_{00}^{\mu_{00}(g_1,g_2^\pi)} p_{10}^{\mu_{10}(g_1,g_2^\pi)} p_{01}^{\mu_{01}(g_1,g_2^\pi)}. \tag{A.8}$$

In (A.8), we define

$$\begin{aligned}
\mu_{11}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i,j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
& + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
\mu_{10}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i,j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} \\
& + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} \\
\mu_{01}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i,j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
& + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
\mu_{00}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i,j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}} \\
& + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}}.
\end{aligned}$$

where $\mathcal{V}^{u'} \triangleq \mathcal{V}_{\mathrm{u}} \setminus \mathcal{V}^s$ is the set of unmatched users vertices and $\mathcal{V}^s$ is the set of seed vertices. Notice that the term summing seed-seed edges is always the same for every $\pi \in \mathcal{S}_{\mathrm{u}}$ since we only permute user vertices. Here, we define

$$\begin{aligned}
\mu'_{11}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
\mu'_{10}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\nsim} j\right\}} \\
\mu'_{01}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\sim} j\right\}} \\
\mu'_{00}(g_1,g_2^\pi) \triangleq & \sum_{i,j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{u'},j\in\mathcal{V}^{s}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}} + \sum_{i\in\mathcal{V}^{s},j\in\mathcal{V}^{u'}} \mathbb{1}_{\left\{i \overset{g_1}{\nsim} j, i \overset{g_2^\pi}{\nsim} j\right\}}.
\end{aligned}$$

We therefore have

$$\mathsf{P}(G_1 = g_1, G_2 = g_2^\pi) \propto p_{11}^{\mu'_{11}(g_1,g_2^\pi)} p_{00}^{\mu'_{00}(g_1,g_2^\pi)} p_{10}^{\mu'_{10}(g_1,g_2^\pi)} p_{01}^{\mu'_{01}(g_1,g_2^\pi)} \tag{A.9}$$

So far the MAP estimator we derived here is exactly the same as the estimator for attributed graph alignment. Applying Lemma 2, we get

$$\begin{aligned}\hat{\pi}_{\mathrm{MAP}}(g_1, g'_2) &= \operatorname*{argmin}_{\pi \in \mathcal{S}_{\mathrm{u}}} \{\mu'_{10}(g_1, g_2^\pi) + \mu'_{01}(g_1, g_2^\pi)\}, \\ &= \operatorname*{argmin}_{\pi \in \mathcal{S}_n} \{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G'_2)) + \Delta^{\mathrm{a}}(G_1, \pi^{-1}(G'_2))\}.\end{aligned}$$

□

Appendix B

Proof of Corollary 1

Corollary 1 (Simplified achievability)**.** Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ under conditions (2.7) and (2.8). If

$$np_{11} + m\psi_{\mathrm{a}} - \log n \to \infty \tag{2.9}$$

then there exists an algorithm that achieves exact alignment *w.h.p.*
If we further have $m = \Omega((\log n)^3)$, then the above condition (2.9) becomes

$$np_{11} + mq_{11} - \log n \to \infty, \tag{2.10}$$

If $m = o((\log n)^3)$, the above condition (2.9) becomes the following.

$$np_{11} + mq_{11} - ma_n - \log n \to \infty, \tag{2.11}$$

where $a_n = q_{11} - \psi_{\mathrm{a}} = O(q_{11}^{3/2})$.

Proof of Corollary 1. In this proof, we first show that, under the assumptions on the user-user edges in condition (2.7) and (2.8), the achievability result becomes

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1)$$

Next, we apply the assumptions on the user-attribute edges from (2.7)(2.8) and derive the two cases in this Corollary by approximating ψ_{a}.

For the user-user edge part, we check the two regimes $p_{11} = \omega(\frac{\log n}{n})$ and $p_{11} = O(\frac{\log n}{n})$ separately. If $p_{11} = \omega(\frac{\log n}{n})$, then with the assumption on the user-user edge density (2.7), we also have $\psi_{\mathrm{u}} = \omega(\frac{\log n}{n})$ because $\psi_{\mathrm{u}} = \Theta(p_{11})$ (see Fact 4). Therefore exact alignment is achievable according to Theorem 1: $\frac{n\psi_{\mathrm{u}}}{2} + m\psi_{\mathrm{a}} - \log n = \omega(\log n) + m\psi_{\mathrm{a}} - \log n = \omega(1)$. Now we check the case when $p_{11} = O(\frac{\log n}{n})$. Notice that under the assumption on the edge correlation (2.8), we have $p_{10} = O(p_{11})$ and $p_{01} = O(p_{11})$ (see Fact 4). Then it follows that the sparsity constrains in Theorem 2 (2.2) (2.3) (2.4) are all satisfied. Therefore, we just need $np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1)$ to guarantee that exact alignment is achievable. Combining the two case, we come the the

conclusion that, under the assumptions (2.7) (2.8), the achievability result simplifies to $np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1)$.

From the above discussion, we further simplify the achievability results so that we can see how it influenced by only on np_{11} and mq_{11}, which are the two only parameters in the converse bound, and thus show when the achievability and converse are tight (up to $\pm\omega(1)$). From the achievability in last step: $np_{11}+m\psi_{\mathrm{a}}-\log n = \omega(1)$, we then need to determine the difference between $m\psi_{\mathrm{a}}$ and mq_{11}. If at the boundary of the achievable region (i.e, the region where $np_{11} + m\psi_{\mathrm{a}} - \log n$ is converging to infinity at constant order), we have $mq_{11} - m\psi_{\mathrm{a}} \leq \omega(1)$ for every np_{11}, then $np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1)$ implies that $np_{11} + mq_{11} - \log n = \omega(1)$, and thus we get the matching achievability and converse in (2.10); Otherwise we keep the achievability at the form of (2.11) which is another way of saying $np_{11} + m\psi_{\mathrm{a}} - \log n \to \infty$.

To see when the achievability matches the converse, first note that at the boundary of achievable region, we have $m\psi_{\mathrm{a}} = C+\log n-np_{11}$ where $C \to \infty$ at constant order. Therefore, $m\psi_{\mathrm{a}}$ at the boundary region attains maximum when $np_{11} = 0$ where $m\psi_{\mathrm{a}} = \Theta(\log n)$. Further combine the conclusions in Fact 4 $\psi_{\mathrm{a}} = \Theta(q_{11})$ and $mq_{11} - m\psi_{\mathrm{a}} = O(mq_{11}^{3/2})$, we come the to conclusion that the gap between mq_{11} and $m\psi_{\mathrm{a}}$ is of order at most $m(\log n/m)^{3/2}$, and the gap attains maximum when $np_{11} = 0$. Therefore, if $m(\log n/m)^{3/2} \leq \omega(1)$, i.e., $m = \Omega((\log n)^3)$, we have the matching achievability and converse as (2.10) and (2.6); otherwise, we left the achievability as (2.11) where a_n stands for the gap between the achievability and converse which grows to infinity faster than constant.

□

Appendix C

Orbit decomposition

Fact 1. *The generating function $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$ of permutation π can be decomposed into*

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) = \prod_{l \geq 1} \mathcal{A}_l(\boldsymbol{x}, z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{y}, z)^{t_l^{\mathrm{a}}},$$

where t_l^{u} is the number of user-user orbits of size l, t_l^{a} is the number of user-attribute orbits of size l.

Proof. Recall the definition of $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$ for a given π

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) = \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}.$$

According to the cycle decomposition on $\pi^{\mathcal{E}}$, we write $\mathcal{E} = \cup_{i \geq 1} \mathcal{O}_i$, where use $\mathcal{O}_i$ is the ith orbit and there are N orbits in total. Then we have

$$\begin{aligned}
\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) &= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)} \\
&= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} \prod_{e \in \mathcal{E}} z^{\delta_\pi(g_e,h_e)} \boldsymbol{x}^{\boldsymbol{\mu}(g_e,h_e)} \boldsymbol{y}^{\boldsymbol{\nu}(g_e,h_e)} && \text{(C.1)} \\
&= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} \prod_{i=1}^{N} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) && \text{(C.2)} \\
&= \sum_{g_{\mathcal{O}_1} \in \{0,1\}^{\mathcal{O}_1}} \sum_{h_{\mathcal{O}_1} \in \{0,1\}^{\mathcal{O}_1}} \cdots \sum_{h_{\mathcal{O}_N} \in \{0,1\}^{\mathcal{O}_N}} \prod_{i=1}^{N} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) && \text{(C.3)} \\
&= \prod_{i=1}^{N} \left(\sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) \right) && \text{(C.4)} \\
&= \prod_{i=1}^{N} \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, \boldsymbol{y}, z) && \text{(C.5)}
\end{aligned}$$

$$= \prod_{l \geq 1} \mathcal{A}_l(\boldsymbol{x}, z^{w_1})^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{y}, z^{w_2})^{t_l^{\mathrm{a}}}. \tag{C.6}$$

Here we use $g_{\mathcal{E}'}$ to denote a subset of g that contains only vertex pairs in $\mathcal{E}'$ and $h_{\mathcal{E}'}$ to denote a subset of h that contains only vertex pairs in $\mathcal{E}'$, where $\mathcal{E}'$ can be any set of vertex pairs. In (C.1), g_e (resp. h_e) represent a subset of g (resp. h) that contains a single vertex pair e. In (C.2), $g_{\mathcal{O}_i}$ (resp. $h_{\mathcal{O}_i}$) represents the subset of g (resp. h) that contains only vertex pairs in orbit $\mathcal{O}_i$. We define $f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})$ as a function of $g_{\mathcal{O}_i}$ and $h_{\mathcal{O}_i}$ where $f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) = \prod_{e \in \mathcal{O}_i} z^{\delta_\pi(g_e, h_e)} \boldsymbol{x}^{\boldsymbol{\mu}(g_e, h_e)}$ if $\mathcal{O}_i$ only contains user-user pairs, and $f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) = \prod_{e \in \mathcal{O}_i} z^{\delta_\pi(g_e, h_e)} \boldsymbol{y}^{\boldsymbol{\nu}(g_e, h_e)}$ if $\mathcal{O}_i$ only contains user-attribute pairs. Equation (C.3) follows because $\mathcal{O}_i$'s are disjoint and their union is $\mathcal{E}$. Note that $f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})$ only concerns vertex pairs in the cycle $\mathcal{O}_i$ since for $e \in \mathcal{O}_i$ we have $\pi^{\mathcal{E}}(e) \in \mathcal{O}_i$. Then, (C.4) follows because $f(g_{\mathcal{O}_i}, h_{G_i})$'s are independent functions. In (C.5), we use $\mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, \boldsymbol{y}, z)$ to denote the generating function for the orbit $\mathcal{O}_i$ where $\mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, \boldsymbol{y}, z) = \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, z)$ if $\mathcal{O}_i$ contains user-user vertex pairs; $\mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, \boldsymbol{y}, z) = \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{y}, z)$ if $\mathcal{O}_i$ contains user-attribute vertex pairs. To see this equation follows, note that if $\mathcal{O}_i$ contains only user-user vertex pairs, then

$$\begin{aligned}
&\sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) \\
&= \sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \prod_{e \in \mathcal{O}_i} z^{\delta_\pi(g_e, h_e)} \boldsymbol{x}^{\boldsymbol{\mu}(g_e, h_e)} \\
&= \sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} z^{\delta_\pi(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})} \boldsymbol{x}^{\boldsymbol{\mu}(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})} \\
&= \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, z).
\end{aligned}$$

If $\mathcal{O}_i$ contains only user-attribute vertex pairs, then

$$\begin{aligned}
&\sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) \\
&= \sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \prod_{e \in \mathcal{O}_i} z^{\delta_\pi(g_e, h_e)} \boldsymbol{y}^{\boldsymbol{\nu}(g_e, h_e)} \\
&= \sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} z^{\delta_\pi(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})} \boldsymbol{y}^{\boldsymbol{\nu}(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})} \\
&= \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{y}, z).
\end{aligned}$$

In (C.6), we apply the fact that orbits of the same size have the same generating function. □

Appendix D

Relation to subsampling model

Fact 4. *If*

$$\begin{aligned} 1-(p_{11}+p_{10}) &= \Theta(1), \\ 1-(p_{11}+p_{01}) &= \Theta(1), \\ \rho_{\mathrm{u}} &= \Theta(1), \end{aligned}$$

then we have $\psi_{\mathrm{u}} = \Theta(p_{11})$,$\psi_{\mathrm{u}} = p_{11} - \Theta(p_{11}^{3/2})$, $p_{10} = O(p_{11})$ *and* $p_{01} = O(p_{11})$. *Note that the same statement holds if we change to* $\boldsymbol{q}$ *and this can be shown through the same proof.*

Proof. To make the notation compact, we consider the equivalent expression from the subsampling model. We have

$$\begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} = \begin{pmatrix} ps_1s_2 & ps_1(1-s_2) \\ p(1-s_1)s_2 & p(1-s_1)(1-s_2)+1-p \end{pmatrix}.$$

The above three conditions on $\boldsymbol{p}$ can be written as

$$1 - ps_1 = \Theta(1), \tag{D.1}$$

$$1 - ps_2 = \Theta(1), \tag{D.2}$$

$$\rho_{\mathrm{u}} = \frac{(1-p)\sqrt{s_1s_2}}{\sqrt{1-ps_1}\sqrt{1-ps_2}} = \Theta(1). \tag{D.3}$$

Combining the above three conditions, we have $s_1 = \Theta(1)$, $s_2 = \Theta(1)$ and $1-p = \Theta(1)$. Therefore, we can directly get $p_{10} = O(p_{11})$ and $p_{01} = O(p_{11})$.

To see $\psi_{\mathrm{u}} = \Theta(p_{11})$, we write ψ_{u} using parameter from subsampling mode and we have

$$\begin{aligned} \psi_{\mathrm{u}} &= (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \\ &= (\sqrt{p_{11}((1-p)+p(1-s_1)(1-s_2))} \\ &\quad - \sqrt{p^2s_1s_2(1-s_1)(1-s_2)})^2 \end{aligned}$$

$$= (1-p)p_{11}\left(\sqrt{1+\tfrac{p(1-s_1)(1-s_2)}{1-p}} - \sqrt{\tfrac{p(1-s_1)(1-s_2)}{1-p}}\right)^2$$
$$= (1-p)p_{11}\frac{1}{\left(\sqrt{1+\tfrac{p(1-s_1)(1-s_2)}{1-p}} + \sqrt{\tfrac{p(1-s_1)(1-s_2)}{1-p}}\right)^2}. \tag{D.4}$$

In (D.4), we have that $(1-p) = \Theta(1)$ and $\frac{1}{\sqrt{1+\frac{p(1-s_1)(1-s_2)}{1-p}}+\sqrt{\frac{p(1-s_1)(1-s_2)}{1-p}}} = \Theta(1)$. Therefore $\psi_{\mathrm{u}} = \Theta(p_{11})$.

To see $\psi_{\mathrm{u}} = p_{11} - \Theta(p_{11}^{3/2})$, we take

$$\begin{aligned}
\psi_{\mathrm{u}} &= (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \\
&= p_{11}p_{00} + p_{10}p_{01} - 2\sqrt{p_{11}p_{00}p_{10}p_{01}} \\
&= p_{11}((1-p) + p(1-s_1)(1-s_2)) \\
&\quad + p^2 s_1 s_2 (1-s_1)(1-s_2) \\
&\quad - \sqrt{p_{11}^2((1-p) + p(1-s_1)(1-s_2))p(1-s_1)(1-s_2)} \\
&= p_{11} - O(p_{11}^{3/2}),
\end{aligned}$$

where the last step follows from $s_1 = \Theta(1)$ and $s_2 = \Theta(1)$.

□

Appendix E

Proof of a new converse

Other than considering the existence of indistinguishable user pairs, here we extend to analysing a more general error event $\mathcal{E}^*$, which represent the existence of permutations that swap exactly two users and fail the MAP estimator. Recall that the MAP estimator simplifies to find a permutation on the n user vertices such that the weighted Hamming distance of the two graphs is minimized, which is equivalent to find a permutation such that the weighted edge overlap of the two graphs is maximized.

$$\begin{aligned}\hat{\pi}_{\mathrm{MAP}}(G_1, G_2') &= \operatorname*{argmin}_{\pi\in\mathcal{S}_n}\{w_1\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + w_2\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))\} \\ &= \operatorname*{argmax}_{\pi\in\mathcal{S}_n}\{w_1\mu_{11}^{\mathrm{u}}(G_1, \pi^{-1}(G_2')) + w_2\mu_{11}^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))\}\end{aligned}$$

This equivalence comes from the conservation of total number of edges in each graph, and a more detailed argument can be found in the proof of Lemma 2 from the Appendix.

To better understand the above objectives and the error event $\mathcal{E}^*$, we rewrite the the edge overlap of two graph using matrix product. We denote the adjacency matrix on the user part by $A^{\mathrm{u}} \in \{0,1\}^{n\times n}$ and the adjacency matrix on the attribute part by $A^{\mathrm{a}} \in \{0,1\}^{n\times m}$. For a permutation π that swaps two users u and v, we use P_{ij} to denote the corresponding permutation matrix. Then we can represent the weighted edge overlap difference as

$$\begin{aligned}&w_1\mu_{11}^{\mathrm{u}}(G_1, G_2) + w_2\mu_{11}^{\mathrm{a}}(G_1, G_2) - w_1\mu_{11}^{\mathrm{u}}(G_1, \pi(G_2)) - w_2\mu_{11}^{\mathrm{a}}(G_1, \pi(G_2)) \\ &= w_1(\langle A_1^{\mathrm{u}}, A_2^{\mathrm{u}}\rangle - \langle A_1^{\mathrm{u}}, P_{ij}A_2^{\mathrm{u}}P_{ij}\rangle) + w_2(\langle A_1^{\mathrm{u}}, P_{ij}A_2^{\mathrm{u}}\rangle - \langle A_1^{\mathrm{u}}, P_{ij}A_2^{\mathrm{u}}\rangle) \\ &= w_1\sum_{k\neq i,j}((A_1^{\mathrm{u}})_{ik} - (A_1^{\mathrm{u}})_{jk})((A_2^{\mathrm{u}})_{ik} - (A_2^{\mathrm{u}})_{jk})) \\ &+ w_2\sum_{k\in\mathcal{V}_{\mathrm{a}}}((A_1^{\mathrm{a}})_{ik} - (A_1^{\mathrm{a}})_{jk})((A_2^{\mathrm{a}})_{ik} - (A_2^{\mathrm{a}})_{jk})) \\ &= w_1\sum_{k\neq i,j}X_{ij,k} + w_2\sum_{k\in\mathcal{V}_{\mathrm{a}}}Y_{ij,k},\end{aligned}$$

where we defined

$$X_{ij,k} \triangleq ((A_1^{\mathrm{u}})_{ik} - (A_1^{\mathrm{u}})_{jk})((A_2^{\mathrm{u}})_{ik} - (A_2^{\mathrm{u}})_{jk}),$$

$$Y_{ij,k} \triangleq ((A_1^{\mathrm{a}})_{ik} - (A_1^{\mathrm{a}})_{jk})((A_2^{\mathrm{a}})_{ik} - (A_2^{\mathrm{a}})_{jk}),$$

Here $X_{ij,k}$ and $Y_{ij,k}$ are discrete random variables taking values from $\{-1, 0, 1\}$. We have $X_{ij,k} = 1$ with probability $2p_{11}p_{00}$, $X_{ij,k} = -1$ with probability $2p_{10}p_{01}$, and $X_{ij,k} = 0$ with probability $1 - 2p_{11}p_{00} - 2p_{10}p_{01}$; $Y_{ij,k} = 1$ with probability $2q_{11}q_{00}$, $Y_{ij,k} = -1$ with probability $2q_{10}q_{01}$, and $Y_{ij,k} = 0$ with probability $1 - 2q_{11}q_{00} - 2q_{10}q_{01}$. For convenience, we define $a \triangleq 2p_{11}p_{00}$, $a' \triangleq 2q_{11}q_{00}$, $b \triangleq 2p_{10}p_{01}$ and $b' \triangleq 2q_{10}q_{01}$.

Correspondingly, the error event

$$\mathcal{E}^* = \{\exists i, j \in \mathcal{V}_{\mathrm{u}}, s.t. w_1 \sum_{k \neq ij} X_{ij,k} + w_2 \sum_{k \in \mathcal{V}_{\mathrm{a}}} Y_{ij,k} \leq 0\}.$$

In the following part of this section, we prove the converse statement by showing the probability of subsets of $\mathcal{E}^*$ is at least a constant. More specifically,

1. In Theorem 3, we have already show with high probability the error event $\mathcal{E}_1^* = \{$there exist indistinguishable user pairs in the intersection graph$\}$ happens.
2. In Lemma 7, we show that with high probability the error event $\mathcal{E}_2^* = \{$there exist $i, j \in \mathcal{V}_{\mathrm{u}}$ such that for all $k \neq i, j$ $X_{ij,k} \leq 0$ and $Y_{ij,k} \leq 0\}$ happens.

Lemma 7. *Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}, m, \boldsymbol{q})$. If there exists a constant ϵ, such that*

$$na + ma' \leq (2 - \epsilon) \log n, \tag{E.1}$$

then $\mathsf{P}(\mathcal{E}_2^*) = 1 - o(1)$

Proof of Lemma 7. In this proof, we show the existence of $i, j \in \mathcal{V}_{\mathrm{u}}$ such that for all $k \neq i, j$ $X_{ij,k} \leq 0$ and $Y_{ij,k} \leq 0$. To this end, for a pair of users (i, j), we define the event $A_{ij} = \{\forall k \in \mathcal{V}_{\mathrm{u}} \setminus \{i, j\}, X_{ij,k} \leq 0,$ and $\forall k \in \mathcal{V}_{\mathrm{a}}, Y_{ij,k} \leq 0\}$ and we further use N to represent the total number of user pairs satisfy the description, i.e., $N \triangleq \sum_{i,j \in \mathcal{V}_{\mathrm{a}}} \mathbb{1}_{\{A_{ij}\}} = \sum_{i,j \in \mathcal{V}_{\mathrm{a}}} N_{ij}$. In the following, we will prove $\mathsf{P}(N > 0) = 1 - o(1)$ using the second moment method.

$$\mathsf{P}(N > 0) \geq \frac{(\mathsf{E}[N])^2}{\mathsf{E}[N^2]}. \tag{E.2}$$

To compute the first and second moments in this upper bound (E.2), we first recall our definition of $X_{ij,k}$ and $Y_{ij,k}$,

$$X_{ij,k} = \begin{cases} 1 & \text{w.p. } a \\ -1 & \text{w.p. } b \\ 0 & \text{w.p. } 1-a-b, \end{cases} \qquad Y_{ij,k} = \begin{cases} 1 & \text{w.p. } a' \\ -1 & \text{w.p. } b' \\ 0 & \text{w.p. } 1-a'-b'. \end{cases}$$

Then, we have the first moment term in equation (E.2)

$$\begin{aligned} \mathsf{E}[N] &= \binom{n}{2}\mathsf{P}(A_{ij}) \\ &= \binom{n}{2}\mathsf{P}(\forall k \neq i,j, X_{ij,k} \leq 0, \text{ or } Y_{ij,k} \leq 0) \\ &= \binom{n}{2} \prod_{k\neq i,j} \mathsf{P}(X_{ij,k} \leq 0) \prod_{k\in\mathcal{V}_{\mathrm{a}}} \mathsf{P}(Y_{ij,k} \leq 0) \qquad \text{(E.3)} \\ &= \binom{n}{2}(1-a)^{n-2}(1-a')^{m} \qquad \text{(E.4)} \\ &\geq \exp\{2\log n - (na+ma') + o(na+ma')\} \\ &\geq \exp\{\epsilon \log n + o(1)\} \\ &= n^{\epsilon+o(1)}. \qquad \text{(E.5)} \end{aligned}$$

Here, equation (E.3) follows because only $X_{ij,k}$ (or $Y_{ij,k}$) is determined by the edges between ik and jk, and thus $X_{ij,k}$'s and $Y_{ij,k}$'s are mutually independent Equation (E.4) follows from plugging in $\mathsf{P}(X_{ij,k} \leq 0) = 1-a$ and $\mathsf{P}(Y_{ij,k} \leq 0) = 1-a'$.

For the second moment term in equation (E.2), we can write is as

$$\begin{aligned} \mathsf{E}[N^2] &= \mathsf{E}\left[\left(\sum_{i,j\in\mathcal{V}_{\mathrm{u}}} N_{ij}\right)^2\right] \\ &= \mathsf{E}\left[\sum_{i<j} N_{ij} + \sum_{\substack{i,j,k,l:\, i<j,k<l \\ i,j,k,l \text{ are distinct}}} N_{ij}N_{kl} + \sum_{\substack{i,j,k,l:\, i<j,k<l \\ \{i,j\} \text{ and } \{k,l\} \text{ share one element}}} N_{ij}N_{kl}\right] \\ &= \mathsf{E}[N] + \binom{n}{2}\binom{n-2}{2}\mathsf{P}(N_{ij}=1, N_{kl}=1) + 6\binom{n}{3}\mathsf{P}(N_{ij}=1, N_{ik}=1). \end{aligned} \qquad \text{(E.6)}$$

Here in the last equation, i,j,k,l represent distinct user vertices. Plugging (E.6) into lower bound (E.2) on the error event from the second moment

method, we have the following.

$$\frac{(\mathsf{E}[N])^2}{\mathsf{E}[N^2]} = \frac{1}{\frac{1}{\mathsf{E}[N]} + \frac{6\binom{n}{3}\mathsf{P}(A_{ij},A_{ik})}{\left(\binom{n}{2}\mathsf{P}(A_{ij})\right)^2} + \frac{\binom{n}{2}\binom{n-2}{2}\mathsf{P}(A_{ij},A_{kl})}{\left(\binom{n}{2}\mathsf{P}(A_{ij})\right)^2}}. \tag{E.7}$$

We then show that the three terms in the denominator of (E.7) are all at most a constant separately.

(1) For the first term in the denominator of (E.7), we have $\frac{1}{\mathsf{E}[N]} = o(1)$, which follows from the lower bound on the first moment (E.5).

(2) For the second term in the denominator of (E.7), we have

$$\begin{aligned}
&\mathsf{P}(N_{ij}=1, N_{ik}=1) = \mathsf{P}(N_{kl}=1)\mathsf{P}(N_{ij}=1|N_{ik}=1) \\
&= \mathsf{P}(A_{kl})\mathsf{P}(A_{ij}|A_{ik}) \\
&= \mathsf{P}(A_{kl}) \prod_{u\in\mathcal{V}_{\mathrm{u}}\setminus\{i,j\}} \mathsf{P}(X_{ij,u}\le 0|X_{ik,u}\le 0) \prod_{u\in\mathcal{V}_{\mathrm{a}}} \mathsf{P}(Y_{ij,u}\le 0|Y_{ik,u}\le 0) \qquad \text{(E.8)} \\
&\le (1-a)^{n-2}(1-a')^m \left(1-\frac{1}{2}a\right)^{n-2}\left(1-\frac{1}{2}a'\right)^m. \qquad \text{(E.9)}
\end{aligned}$$

Here (E.8) holds because conditioned on A_{ik}, $X_{ij,u}$ and $Y_{ij,u}$ is only function of the edges between j and u, thus $X_{ij,u}$'s and $Y_{ij,u}$'s are mutually independent. In equation (E.9), we plug in $\mathsf{P}(A_{kl}) = (1-a)^{n-2}(1-a')^m$ and we use the following upper bounds on $\mathsf{P}(X_{ij,u}\le 0|X_{ik,u}\le 0)$ and $\mathsf{P}(Y_{ij,u}\le 0|Y_{ik,u}\le 0)$. We have

$$\begin{aligned}
&\mathsf{P}(X_{ij,u}\le 0|X_{ik,u}\le 0) \\
&= \frac{p_{11}(1-p_{00}^2+p_{01}+p_{10}+p_{00}(1-p_{11})^2)}{1-2p_{00}p_{11}} \\
&\le \frac{1-3p_{11}p_{00}}{1-2p_{11}p_{00}} \\
&\le 1-p_{11}p_{00} = 1-\frac{1}{2}a
\end{aligned}$$

and

$$\begin{aligned}
&\mathsf{P}(Y_{ij,u}\le 0|Y_{ik,u}\le 0) \\
&= \frac{q_{11}(1-q_{00}^2+q_{01}+q_{10}+q_{00}(1-q_{11})^2)}{1-2q_{00}q_{11}} \\
&\le \frac{1-3q_{11}q_{00}}{1-2q_{11}q_{00}}
\end{aligned}$$

$$\leq 1 - q_{11}q_{00} = 1 - \frac{1}{2}a.$$

Therefore, we get the following.

$$\begin{aligned}
&\frac{6\binom{n}{3}\mathsf{P}(A_{ij}, A_{ik})}{\left(\binom{n}{2}\mathsf{P}(A_{ij})\right)^2} \\
&= \Theta\Bigg(\exp\Bigg\{-n\left(\log(1-a) - \log\left(1-\frac{a}{2}\right)\right) \qquad \text{(E.10)} \\
&\qquad\qquad - m\left(\log(1-a') - \log\left(1-\frac{a'}{2}\right)\right) - \log n\Bigg\}\Bigg) \\
&= \Theta\left(\exp\left\{-\log n + \frac{na}{2} + \frac{ma'}{2} + o(\frac{na}{2} + \frac{ma'}{2})\right\}\right) \\
&\leq \exp(-\epsilon \log n + o(\log n)) \\
&= n^{-\Theta(1)}. \qquad \text{(E.11)}
\end{aligned}$$

(3) For the last term in the the denominator of (E.7) , we have

$$\begin{aligned}
&\mathsf{P}(N_{ij} = 1, N_{kl} = 1) = \mathsf{P}(N_{ij} = 1 | N_{kl} = 1)\mathsf{P}(N_{kl} = 1) \\
&= \mathsf{P}(A_{ij}|A_{kl})\mathsf{P}(A_{kl}) \\
&\leq \mathsf{P}(A_{kl})\mathsf{P}(\forall u \neq ijkl, X_{ij,u} \leq 0 \text{ or } Y_{ij,u} \leq 0 | A_{kl}) \\
&= \mathsf{P}(A_{kl}) \prod_{u \neq ijkl, u \in \mathcal{V}_{\mathrm{u}}} \mathsf{P}(X_{ij,u} \leq 0) \prod_{u \in \mathcal{V}_{\mathrm{a}}} \mathsf{P}(Y_{ij,u} \leq 0) \qquad \text{(E.12)} \\
&= (1-a)^{2n-6}(1-a')^{2m}. \qquad \text{(E.13)}
\end{aligned}$$

Equation (E.12) follows since $X_{ij,u}$'s (or $Y_{ij,k}$'s) are mutually independent and they are independent of A_{kl}. In (E.9), we plug in $\mathsf{P}(A_{kl}) = (1-a)^{n-2}(1-a')^m$, $\mathsf{P}(X_{ij,u} \leq 0) = 1-a$, and $\mathsf{P}(Y_{ij,u} \leq 0) = 1-a'$. Therefore, we have

$$\frac{\binom{n}{2}\binom{n-2}{2}\mathsf{P}(A_{ij}, A_{kl})}{\left(\binom{n}{2}\mathsf{P}(A_{ij})\right)^2} = \Theta\left((1-a)^{-2}\right) = \Theta(1). \qquad \text{(E.14)}$$

Plugging the three terms (E.5), (E.11)and (E.14) into (E.6), we have

$$\frac{(\mathsf{E}[N])^2}{\mathsf{E}[N^2]} = \Theta(1). \qquad \text{(E.15)}$$

Therefore, we show that the error probability is not diminishing using to the second moment method

$$\mathsf{P}(N > 0) \geq \frac{(\mathsf{E}[N])^2}{\mathsf{E}[N^2]} = \Theta(1).$$

□