# Attributed Graph Alignment

Ning Zhang, Ziao Wang, *Graduate Student Member, IEEE*, Weina Wang, *Member, IEEE*, and Lele Wang, *Member, IEEE*

***Abstract*— Motivated by various data science applications including de-anonymizing user identities in social networks, we consider the graph alignment problem, where the goal is to identify the vertex/user correspondence between two correlated graphs. Existing work mostly recovers the correspondence by exploiting the user-user connections. However, in many real-world applications, additional information about the users, such as user profiles, might be publicly available. In this paper, we introduce the attributed graph alignment problem, where additional user information, 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 certain parameter regimes. Our results span the full spectrum between models that only consider user-user connections and models where only attribute information is available.**

***Index Terms*— Graph theory, statistics, inference algorithms.**

## I. Introduction

THE graph alignment problem, also known as graph matching problem or noisy graph isomorphism problem, has received increasing attention in recent years, brought into prominence by applications in a wide range of areas [1], [2], [3]. For instance, in social network deanonymization [4], [5], two graphs are given, 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 the 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–Rényigraph pair* model (see, e.g., [6], [7], [8]). 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 [6], [7], [8], [9]. The best known information-theoretic limits are proved in [8] and [9], 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 Narayanan and Shmatikov [10]. 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 paper, 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). Here, the user vertices represent the entities that need to be aligned; while the attribute vertices

Manuscript received 16 March 2023; revised 10 March 2024; accepted 10 May 2024. Date of publication 21 May 2024; date of current version 16 July 2024. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery under Grant RGPIN-2019-05448, in part by the NSERC Collaborative Research and Development under Grant CRDPJ 54367619, and in part by the U.S. National Science Foundation (NSF) under Grant CNS-2007733 and Grant ECCS-2145713. An earlier version of this paper was presented in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9517967]. *(Corresponding author: Ziao Wang.)*

Ning Zhang was with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada. She is now with the Department of Statistics, University of Oxford, OX1 3LB Oxford, U.K. (e-mail: ning.zhang@stats.ox.ac.uk).

Ziao Wang and Lele Wang are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: ziaow@ece.ubc.ca; lelewang@ece.ubc.ca).

Weina Wang is with the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: weinaw@cs.cmu.edu).

Communicated by L. Lai, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3403810.

Digital Object Identifier 10.1109/TIT.2024.3403810

0018-9448 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

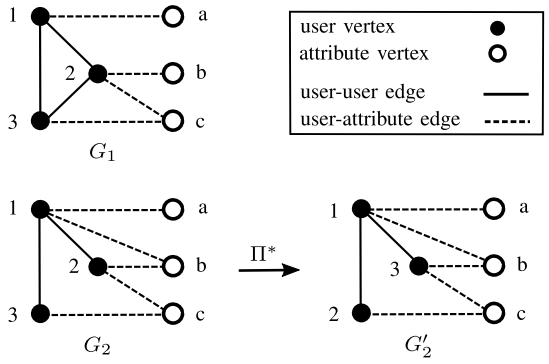


Fig. 1. Example of attributed Erdős–Rényi graph pair: Graph $G_1$ and $G_2$ are generated on the same set of vertices. Anonymized graph $G_2'$ is obtained through applying $\Pi^* = (1)(2,3)$ only on $\mathcal{V}_a$ of $G_2$ (permutation $\Pi^*$ is written in cycle notation).

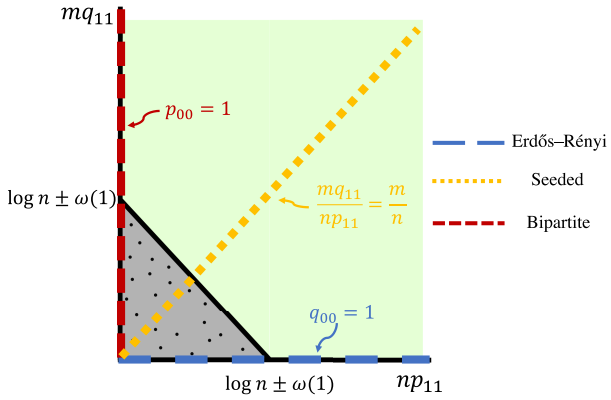


Fig. 2. The green region in the figure is information theoretically achievable and the shaded grey region is not achievable. The three lines in the figure represent three specialized settings: the blue line (correlated Erdős–Rényi model) is obtained by setting $q_{00} = 1$; the yellow line (seeded Erdős–Rényi model) is obtained by setting $\boldsymbol{p} = \boldsymbol{q}$; the red line (correlated bipartite model) is obtained by setting $p_{00} = 1$. Their intersections with the achievable and non-achievable region give the information-theoretic limits of the correlated Erdős–Rényi model, seeded Erdős–Rényi model and the correlated bipartite model separately.

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.

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 restricting the graph parameters to a certain regime, while deferring the general result to Section III. In this regime we assume that the correlation coefficient of the user-user edges is at least $\Omega(\frac{(\log n)^2}{\sqrt{n}})$ and correlation coefficient of the user-attribute edges is at least $\Omega(\frac{(\log n)^{3/2}}{\sqrt{m}})$. Together with two other conditions on the edge sparsity, we establish the following asymptotically matching achievability and converse results as $n \to \infty$ (See Corollary 1 for the formal statement).

- If $np_{11} + mq_{11} - \log n \to \infty$, then there exists an algorithm that achieves exact alignment with high probability (w.h.p.).
- If $np_{11} + mq_{11} - \log n \to -\infty$, then no algorithm guarantees exact alignment w.h.p.

The achievability and converse results are illustrated in Figure 2. Here, $np_{11}$ is the average number of common users between $G_1$ and $G_2$ that are connected to an 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 noting that the average common vertex degree in attribute $mq_{11}$ highlights the extra benefit from attribute information, compared to the achievability result $np_{11} - \log n \to \infty$ when the attribute is not available.

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 2). Below we highlight how the specializations of our results compare with the existing graph alignment literature under three specialized settings. The detailed comparison is given in Section IV.

- 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.
  Our specialized results recover the information-theoretic limits on Erdős–Rényi graph alignment in [9].
- Specializing our model by setting $\boldsymbol{p} = \boldsymbol{q}$, we can then treat the $m$ attribute vertices as pre-aligned user vertices and obtain the seeded Erdős–Rényi model. Our specialized result reveals certain achievable and converse region that is unknown in the literatures [9], [11], and [12].
- Specializing our model by setting $p_{00} = 1$, we remove all of the user-user edges and obtain the correlated bipartite graph pair model. Our specialized achievability result

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

recovers the best-known achievable region for bipartite alignment in the literature [13], and reveals certain converse region that is unknown in [13].

The main contributions of this paper are summarized as follows.

1) ***Model Formulation.*** We propose the attributed Erdős–Rényi pair model, which incorporates both the graph topology similarity 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 setting in the graph alignment literature and includes several popular models as its special cases.
2) ***Information theoretic limits.*** We establish achievability and converse results on exactly aligning random attributed graphs, where the conditions are tight under certain parameter regimes.
   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.
   When specialized to the seeded graph alignment and bipartite graph alignment models, our result reveals certain achievable and converse region that is unknown in the literature.
3) ***Proof techniques.*** The proof techniques for the achievability results are mainly inspired by the previous study on Erdős–Rényi graph alignment [9]. For the converse results, we study the phase-transition phenomenon on the existence of *indistinguishable vertex pairs*, which may be of independent interest.

### A. Related Work

The exact graph alignment problem has been studied under various random graph models. One of the most popular random graph 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 graphs achieve the maximum edge overlap. While the optimal strategy requires exponential time complexity, numerous studies have proposed polynomial-time approaches that exactly solve the graph alignment problem with high probability [14], [15], [16], [17], [18].

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 [8] and [9] by analyzing error event of the MAP estimator. In [9], the authors prove achievability in the regime $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq (2+\epsilon)\log n$. Under certain sparsity conditions, they also show that the achievable region can be improved to $np_{11} \geq \log n + \omega(1)$. In [8], the authors consider a special case of the Erdős–Rényi graph pair model called *symmetric subsampling model*. In this model, it is assumed that

$$p_{11} = ps^2,\ p_{01} = p_{10} = ps(1-s),\ p_{00} = 1 - 2ps + ps^2 \tag{1}$$

for some $p, s \in [0,1]$. Under this model, the authors prove the achievability in the regime $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \geq (1+\epsilon)\log n$. For the converse, [9] proves that the permutation cannot be exactly recovered with high probability if $np_{11} \leq \log n - \omega(1)$ by showing the existence of isolated vertices in the intersection graph $G_1 \wedge G_2$. Under the general Erdős–Rényi graph pair model, [8] shows the impossibility of exactly recovering the permutation in the regime $n(\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \leq (1-\epsilon)\log n$ by showing the existence of permutations that fails the MAP estimator by swapping two vertices. To summarize the aforementioned results, matching achievability and converse for exact recovery is derived under certain sparsity assumptions in [9], and for the special case of the symmetric subsampling model, [8] provides almost tight achievability and converse bounds, with a gap of width $2\epsilon \log n$ between the established bounds. Closing the gap for the general Erdős–Rényi graph pair model is still an open problem.

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 [11], [19], [20]. Some other more general settings treat any form of side information as vertex attributes and formulate this as the attribute graph alignment problem [21].

There is a line of empirical studies on the attributed graph alignment [21], [22], [23], yet, to the best of our knowledge, there is no known result on information-theoretic limits on graph alignment with attribute information.

## II. Model

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.

### A. 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 $[n] \triangleq \{1, 2, 3, \ldots, n\}$. Assume that the attribute vertex set $\mathcal{V}_{\mathrm{a}}$ consists of $m$ vertices, and $m$ scales as a function of $n$.

### B. 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

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

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

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } p_{11}, \\ (1,0) & \text{w.p. } p_{10}, \\ (0,1) & \text{w.p. } p_{01}, \\ (0,0) & \text{w.p. } p_{00}, \end{cases} \tag{2}$$

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

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } q_{11}, \\ (1,0) & \text{w.p. } q_{10}, \\ (0,1) & \text{w.p. } q_{01}, \\ (0,0) & \text{w.p. } q_{00}, \end{cases} \tag{3}$$

where $q_{11}, q_{10}, q_{01}, q_{00}$ are probabilities that 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, recall that there are no edges between attribute vertices in our model.

For compactness of notation, we represent the joint distributions in (2) and (3) 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 [6].

### C. 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 $\Pi^*$ on the user vertex set of $G_2$, where the permutation $\Pi^*$ is unknown. More explicitly, each user vertex $i$ in $G_2$ is re-labeled as $\Pi^*(i)$ in $G_2'$. The permutation $\Pi^*$ 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 $\Pi^*$. 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 $\Pi^*$. 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 exact alignment is achievable *with high probability (w.h.p.)* if there exists $\hat{\pi}$ such that $\lim_{n\to\infty} \mathsf{P}(\hat{\pi}(G_1, G_2') \neq \Pi^*) = 0$.

### D. Reminder of the Landau Notation

| Notation | Definition |
|---|---|
| $f(n) = \omega(g(n))$ | $\lim_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} = \infty$ |
| $f(n) = o(g(n))$ | $\lim_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} = 0$ |
| $f(n) = O(g(n))$ | $\limsup_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} < \infty$ |
| $f(n) = \Omega(g(n))$ | $\liminf_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} > 0$ |
| $f(n) = \Theta(g(n))$ | $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ |

## III. Main Results

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

Throughout the remainder of the paper, we define

$$\psi_{\mathrm{u}} \triangleq (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \tag{4}$$

$$\psi_{\mathrm{a}} \triangleq (\sqrt{q_{11}q_{00}} - \sqrt{q_{10}q_{01}})^2. \tag{5}$$

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

$$\tfrac{1}{2}n\psi_{\mathrm{u}} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{6}$$

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

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{7}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{8}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{9}$$

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{10}$$

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

$$-n\log(1 - 2p_{11} + 2p_{11}^2) - m\log(1 - 2q_{11} + 2q_{11}^2) - 2\log n \to -\infty, \tag{11}$$

then for any estimator, the probability of error is bounded away from zero.

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 certain

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

conditions on the sparsity and correlation of the two graphs. To make the notation compact, we consider the equivalent expression from the subsampling model, where

$$\begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} = \begin{pmatrix} ps_{\mathrm{u},1}s_{\mathrm{u},2} & ps_{\mathrm{u},1}(1-s_{\mathrm{u},2}) \\ p(1-s_{\mathrm{u},1})s_{\mathrm{u},2} & p(1-s_{\mathrm{u},1})(1-s_{\mathrm{u},2})+1-p \end{pmatrix},$$

and

$$\begin{pmatrix} q_{11} & q_{10} \\ q_{01} & q_{00} \end{pmatrix} = \begin{pmatrix} qs_{\mathrm{a},1}s_{\mathrm{a},2} & qs_{\mathrm{a},1}(1-s_{\mathrm{a},2}) \\ q(1-s_{\mathrm{a},1})s_{\mathrm{a},2} & q(1-s_{\mathrm{a},1})(1-s_{\mathrm{a},2})+1-q \end{pmatrix}.$$

Under the subsampling model, the generation of $G_1$ and $G_2$ is modelled as a two-step random process. We first generate a base graph $G$, where an edge exists between each user-user pair with probability $p$ and an edge exists between each user-attribute pair with probability $q$. To generate graph $G_1$, each user-user edge in $G$ is kept with probability $s_{\mathrm{u},1}$ and each user-attribute edge in $G$ is kept with probability $s_{\mathrm{a},1}$. Similarly $G_2$ is generated by keeping each user-user edge in $G$ with probability $s_{\mathrm{u},2}$ and each user-attribute edge in $G$ with probability $s_{\mathrm{a},2}$. A random permutation is then applied on the user of $G_2$ to generate $G_2'$. As a mild restriction on the sparsity of the graphs, we assume that the base graph edge probabilities $p$ and $q$ are not going to 1, i.e.,

$$1-p=\Theta(1), \tag{12}$$
$$1-q=\Theta(1). \tag{13}$$

Moreover, we assume the following bounds on the vanishing speed of subsampling probabilities $s_{\mathrm{u},1}, s_{\mathrm{u},2}, s_{\mathrm{a},1}$ and $s_{\mathrm{a},2}$

$$s_{\mathrm{u},1}s_{\mathrm{u},2}=\Omega\left(\frac{(\log n)^4}{n}\right), \tag{14}$$
$$s_{\mathrm{a},1}s_{\mathrm{a},2}=\Omega\left(\frac{(\log n)^3}{m}\right). \tag{15}$$

*Corollary 1 (Simplified Achievability):* Consider the attributed Erdős–Rényi pair $\mathcal{G}(n,\boldsymbol{p};m,\boldsymbol{q})$. Under conditions (12)-(15), we have tight achievability and converse for exact recovery. That is if

$$np_{11}+mq_{11}-\log n\to\infty, \tag{16}$$

then the MAP estimator achieves exact alignment *w.h.p*, and if

$$np_{11}+mq_{11}-\log n\to-\infty, \tag{17}$$

then the error probability of any estimator is bounded away from zero.

The proof of Corollary 1 can be found in Appendix F. We visualize the matching achievability and converse results under conditions (12)-(15) in Figure 2. How to close the gap in the case where at least one of (12)-(15) is not satisfied is an open problem.

## IV. Comparison

In this section, we specialize our main results (Theorems 1, 2, and 3) on exact alignment of the attributed Erdős–Rényi pair model to three closely related graph alignment problems: the Erdős–Rényi graph alignment, the seeded Erdős–Rényi graph alignment, and the bipartite graph alignment. We compare them with the best-known results in the literature. The main purpose of this comparison is to illustrate that our general results on the attributed graph alignment problem can recover most of the best-known existing results on these three specialized problems, and improve the state-of-the-art in certain cases. While it is possible that the best-known results can be further improved to get sharper bounds, this refinement is not the main focus of our comparison.

### A. Erdős–Rényi Graph Pair

The correlated Erdős–Rényi pair model $\mathcal{G}(n,\boldsymbol{p})$ is the setting most commonly studied for graph alignment tasks that consider only graph topology similarity [6], [7], [8], [9], [14]. This model generates graph pairs that contain only user vertices. 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 generated jointly in the following way: for a pair of users $e\in\binom{\mathcal{V}_{\mathrm{u}}}{2}$, we have

$$(G_1,G_2)(e)=\begin{cases}(1,1) & \text{w.p. } p_{11},\\ (1,0) & \text{w.p. } p_{10},\\ (0,1) & \text{w.p. } p_{01},\\ (0,0) & \text{w.p. } p_{00}.\end{cases} \tag{18}$$

The anonymized graph $G_2'$ is obtained by applying a random permutation $\Pi^*$ on the vertices of $G_2$. This model can be specialized from the attributed graph pair model by setting the number of attributes $m=0$ or $q_{00}=1$. For aligning the correlated Erdős–Rényi pair, the best-known information-theoretic limits are established in [8] and [9] and we state the combined results here for ease of comparison.

*Theorem 4 (Best-Known Information Theoretic Limits [8], [9]):* Consider the correlated Erdős–Rényi pair $\mathcal{G}(n,\boldsymbol{p})$.

*Achievability* If

$$n\psi_{\mathrm{u}}\geq 2\log n+\omega(1), \tag{19}$$

or

$$p_{11}=O\left(\tfrac{1}{\log n}\right), \tag{20}$$
$$p_{10}+p_{01}=O\left(\tfrac{1}{\log n}\right), \tag{21}$$
$$\frac{p_{10}p_{01}}{p_{11}p_{00}}=O\left(\tfrac{1}{(\log n)^3}\right), \tag{22}$$
$$np_{11}=\log n+\omega(1), \tag{23}$$

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

*Converse* If there exist a constant $\epsilon\in(0,1)$ such that

$$n\psi_{\mathrm{u}}\leq(1-\epsilon)\log n, \tag{24}$$

or

$$np_{11}\leq\log n-\omega(1), \tag{25}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

then for any estimator, the probability of error is bounded away from zero.

*Remark 1:* We point out that in the dense regime, i.e., at least one of conditions (20), (21) and (22) is not satisfied, Theorem 4 in [8] provides a tighter achievability result. However, as we mentioned in the introduction, the result is limited to the symmetric subsampling model (1). In this section, we focus on the comparison under the *general* Erdős–Rényi pair model, so the result from Theorem 4 in [8] is not listed as one of the best known information theoretic limit. On the other hand, the converse result in [8] is not limited to the symmetric subsampling model. Thus, we include the result as equation (24) in Theorem 4.

We now specialize the attributed Erdős–Rényi pair model to the correlated Erdős–Rényi pair by setting $q_{00} = 1$. Theorems 1, 2, and 3 simplify to the following.

*Theorem 5 (Specialization from attributed Erdős–Rényi pair):* Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ with $q_{00} = 1$.

*Achievability:* If

$$n\psi_{\mathrm{u}} \geq 2\log n + \omega(1), \tag{26}$$

or

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{27}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{28}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{29}$$

$$np_{11} = \log n + \omega(1), \tag{30}$$

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

*Converse:* If

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

then for any estimator, the probability of error is bounded away from zero.

*Remark 2:* When specialized to the Erdős–Rényi pair model, our achievability result recovers the best-known achievability result from [8] and [9], while our converse result is a strict subset of that given by conditions (24) and (25). To see the achievability results in Theorems 4 and 5 are equivalent, we observe that the difference between the region characterized by (20)–(23) and the region characterized by (27)–(30) is given by $p_{11} = \omega(\frac{\log n}{n})$ and $p_{11} = O(\frac{1}{\log n})$. However, under the assumptions (28) and (29), we know that $p_{00} = 1 - o(1)$ and $\frac{p_{10}p_{01}}{p_{00}p_{11}} = o(1)$. If $p_{11} = \omega(\frac{\log n}{n})$, then these further imply that $\psi_{\mathrm{u}} = (1 + o(1))p_{11}p_{00} = \omega(\frac{\log n}{n})$, i.e., the difference between the two regions falls in the achievable region characterized by (26). Thus, the achievability region in Theorem 5 is exactly the same as that in Theorem 4. For the converse, it is an open question whether a converse result for the attributed Erdős–Rényi pair model can be established, which recovers condition (24) when specialized to the Erdős–Rényi pair model.

### B. Seeded Erdős–Rényi Graph 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 a set 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 [5], [11], [19], [24], and [25].[1] Moreover, achievability results on unseeded graph alignment problem also trivially imply achievability results on seeded graph alignment problem. To the best of our knowledge, the best information-theoretic limits of the seeded alignment problem are given by [9], [11], and [12]. For the simplicity of our discussion, we focus only on the symmetric subsampling model (1) with sparsity conditions

$$p = 1 - \Theta(1) \text{ and } s = \Omega\left(\frac{(\log n)^2}{\sqrt{n}}\right). \tag{31}$$

*Theorem 6 (Best-Known Information-Theoretic Limits in the Sparse and Symmetric Regime [9], [11], [12]):* Consider the seeded Erdős–Rényi graph pair $\mathcal{G}(n, m, \boldsymbol{p})$ satisfying conditions (1) and (31).

Achievability from [9]: Assume

$$(n+m)p_{11} = \log(n+m) + \omega(1). \tag{32}$$

Then the unseeded MAP estimator achieves exact alignment w.h.p.

Achievability from [12]: Assume $s = \Theta(1)$ and $p = o(1)$.

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

$$(n+m)p_{11} \geq (1+\epsilon)\log n, \tag{33}$$

then the ATTRRICH algorithm in [12] achieves exact alignment w.h.p.

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

$$np_{11} = \log n + \omega(1), \tag{34}$$

$$mp_{11} \geq \frac{2\log n}{\tau \log(p_{11}/(p_{11}+p_{10})^2)}, \tag{35}$$

then the ATTRSPARSE algorithm in [12] achieves exact alignment w.h.p.

Converse from [11] Consider the seeded Erdős–Rényi graph pair $\mathcal{G}(n, m, \boldsymbol{p})$. If

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

then for any estimator, the probability of error is bounded away from zero.

*Remark 3 (Efficient Algorithms for Seeded Graph Alignment):* We comment that the seeded graph alignment algorithms proposed in [11] and [12] can be implemented polynomial-time, while the unseeded MAP estimator in [9] requires exponential time to implement. Under the seeded graph alignment problem, the best-known feasible range of

[1] In the literature, both random [5] and deterministic [24] seed sets are considered. Here, we focus on the deterministic seed set setting which is closely related to our attributed Erdős–Rényi pair model.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

graph parameters for achieving exact recovery by efficient algorithms is given together by [11] and [12].

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 [11] exploit 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 exact alignment in the seeded Erdős–Rényi pair model $\mathcal{G}(n, m, \boldsymbol{p})$ and the information-theoretic limits on exact alignment in the specialized attributed Erdős–Rényi pair model $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{p})$ are identical.

*Proof:* See Appendix B. □

Based on Lemma 1, we directly obtain the achievability and converse results on seeded graph alignment from Theorems 1, 2, and 3 by setting $\boldsymbol{p} = \boldsymbol{q}$. In the following, we demonstrate that the specialized result reveals certain achievable and converse region for seeded graph alignment that is unknown in the literature.

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

*Achievability:* If

$$(n+m)\psi_{\mathrm{u}} \geq 2\log n + \omega(1), \tag{36}$$

or

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{37}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{38}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{39}$$

$$np_{11} + m\psi_{\mathrm{u}} = \log n + \omega(1), \tag{40}$$

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

*Converse:* If

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

then for any estimator, the probability of error is bounded away from zero.

Moreover, if the two seeded graphs $G_1$ and $G_2'$ satisfy the two assumptions (1) and (31), then we have the following matching achievability and converse results.

*Corollary 2 (Threshold for sparse seeded Erdős–Rényi pair):* Consider the seeded Erdős–Rényi pair model $\mathcal{G}(n, m, \boldsymbol{p})$ under conditions (1) and (31).

*Achievability:* If

$$(n+m)p_{11} \geq \log n + \omega(1), \tag{41}$$

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

*Converse:* If

$$(n+m)p_{11} \leq \log n - \omega(1), \tag{42}$$

then for any estimator, the probability of error is bounded away from zero.

*Proof:* See Appendix G. □

*Remark 4:* In Corollary 2, we obtain asymptotically tight achievability and converse for seeded graph alignment under the symmetric subsampling model satisfying conditions (1) and (31).

*Remark 5 (Comparison between achievability results):* The achievability result in Corollary 2 reveals certain achievable region that is unknown in the literatures [9] and [12]. In the regime of $mp_{11} = \Omega(\log n)$, if $(n+m)p_{11}$ is at least $\log n + \omega(1)$ but less than $\log(n+m) + \omega(1)$ and $(1+\epsilon)\log n$ for any constant $\epsilon$, exact alignment is known to be achievable by Corollary 2, but not by Theorem 6. In the following, we present an example which is in the achievable region of Corollary 2, but not in the achievable region of Theorem 6. Assume that

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

We see that condition (41) holds because $(n+m)p_{11} = \log n + \log\log n = \log n + \omega(1)$. Moreover, condition (31) is satisfied because $s = 1$ in this case. However, condition (32) $(n+m)p_{11} = \log(n+m) + \omega(1)$ in Theorem 6 does not hold because $(n+m)p_{11} < 2\log n < \log(n+m) + \omega(1)$, and condition (33) $(n+m)p_{11} \geq (1+\epsilon)\log n$ in Theorem 6 does not hold because $(n+m)p_{11} = \log n + \log\log n < (1+\epsilon)\log n$ for any positive constant $\epsilon$. So this example lies in the achievable region of Corollary 2, but not in that of Theorem 6.

However, we comment that the improvement mentioned above is natural. In [9], the seedless graph alignment problem is considered. The results in [9] is included in the comparison because its achievable region for seedless graph alignment trivially implies achievable region for seeded graph alignment. When specializing the region in [9] to the scenario of seeded graph alignment, both seed vertices and non-seed vertices are viewed as vertices to align, and hence resulting in the $\log(m+n)$ terms on the right-hand side of (32). However, the identities of the seed vertices are already known, and there are actually just $n$ non-seed vertices to align. This causes the natural improvement to the $\log n$ term on the right-hand side of (41). Moreover, note that this improvement is non-negligible only when $m = \omega(n)$, i.e., the number of seeds greatly surpass the number of non-seeds. The improvement beyond the achievable region in [12] is natural as well. This is because the MAP estimation studied in this work requires exponential computational time, while the algorithms in [12] can be implemented in polynomial time. To make the comparison fair, we also mention that there exists certain known achievable region that is not covered by Corollary 2.

*Remark 6 (Comparison Between Converse Results):* Corollary 2 includes certain converse region in the regime $m = \omega(n)$, while the converse region in Theorem 6 is exclusive to the regime $m = O(n)$. To make the comparison fair, we comment that the converse bound of Theorem 6

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

strictly contains the converse region of Corollary 2 in the regime of $m = O(n)$.

### C. Bipartite Graph Pair

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

$$(G_1, G_2)(e) = \begin{cases} (1,1) & \text{w.p. } q_{11}, \\ (1,0) & \text{w.p. } q_{10}, \\ (0,1) & \text{w.p. } q_{01}, \\ (0,0) & \text{w.p. } q_{00}. \end{cases} \tag{43}$$

The anonymized graph $G_2'$ is obtained by applying a random permutation $\Pi^*$ only on the user vertex set of $G_2$. Aligning such correlated bipartite graph is also known as a special case of the database alignment problem [13]. The best-known information-theoretic limits of database alignment are studied in [13]. We restate the achievability and converse results from [13] when specialized to the case of bipartite graph pair alignment in Theorem 8.

*Theorem 8 (Best-Known Information-Theoretic Limits [13]):* Consider the bipartite graph pair model $\mathcal{G}(n, m, \boldsymbol{q})$.

*Achievability:* If

$$-\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) \geq \log n + \omega(1), \tag{44}$$

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

*Converse:* If

$$-\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) \leq (1 - \Omega(1))\log n, \tag{45}$$

then for any estimator, the probability of error is bounded away from zero.

*Remark 7:* In [13], the left-hand side of both inequalities in Theorem 8 are stated in a different, yet equivalent, form. To state it, we need to introduce two definitions. Let $A = \begin{smallmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{smallmatrix}$ be a $2 \times 2$ probability matrix with all non-negative entries and $A_{00} + A_{01} + A_{10} + A_{11} = 1$. We define $A^{\otimes k}$ to be a $2^k \times 2^k$ probability matrix with rows and columns both indexed by $\{0,1\}^k$, and for $a, b \in \{0,1\}^k$, we have $A^{\otimes k}_{a,b} = \Pi_{i=1}^k A_{a_i,b_i}$. Matrix $A^{\otimes k}$ is called the $k$th *tensor product* of $A$. For a probability matrix $A$ and an integer $l \geq 2$, we define the order-$l$ *cycle mutual information* of $A$ to be $I_l^\circ(A) = \frac{1}{1-l}\log \mathrm{tr}((ZZ^T)^l)$, where $Z$ is a matrix with same dimension as $A$ and $Z_{ij} = \sqrt{A_{ij}}$ for any index pair $(i, j)$. In [13], the left-hand side of both inequalities are given as $\frac{1}{2}I_2^\circ(Q^{\otimes m})$, where $Q = \begin{smallmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{smallmatrix}$. According to [13], the cycle mutual information satisfies a nice property that $I_2^\circ(Q^{\otimes m}) = mI_2^\circ(Q)$. Furthermore, we can calculate the 2-cycle mutual information of $Q$ as $I_2^\circ(Q) = -\log(1 - 2\psi_{\mathrm{a}})$, which implies that $\frac{1}{2}I_2^\circ(Q^{\otimes m}) = -\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}})$.

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., $p_{00} = 1$. Correspondingly, we obtain the following achievable and converse result on bipartite graph alignment from Theorems 1 and 3.

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

*Achievability:* If

$$m\psi_{\mathrm{a}} \geq \log n + \omega(1), \tag{46}$$

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

*Converse:* If

$$mq_{11} \leq \log n - \omega(1), \tag{47}$$

then for any estimator, the probability of error is bounded away from zero.

*Remark 8:* The MAP estimator for bipartite graph alignment can be implemented by solving an assignment problem using the Hungarian method within polynomial time [26].

*Remark 9 (Comparison between achievable regions):* The achievable region in Theorem 9 is a strict subset of the achievable region in Theorem 8. This because $-\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) > m\psi_{\mathrm{a}}$. However, in the derivation steps (69)–(71) of Theorem 1, we were replacing the logarithm term by applying $\log(1 + x) \leq x$. Without this step of loosening the bound, equation (46) can be replaced by $-\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) \geq \log n + \omega(1)$, which is the same as equation (44). Therefore, although the achievability region (46) in Theorem 9 does not directly recover the achievable region (44) in Theorem 8, it can be improved to the achievable region (44) by a slight modification of derivations steps (69)–(71).

*Remark 10 (Comparison Between Converse Regions):* The converse result (47) in Theorem 9 includes certain region that is not covered by the best-known converse region for bipartite alignment (45) in Theorem 8. This new region stems from the difference between the $-\Omega(1)\log n$ term on the right-hand side of (45) and the $-\omega(1)$ term on the right-hand side of (47). To illustrate the improved region, consider an instance of the parameters satisfying $q_{11} = \frac{\log n - \log\log n}{m}$, $q_{00} = 1 - q_{11}$, $q_{01} = q_{10} = 0$ and $m = \omega(\log n)$. Because $mq_{11} = \log n - \log\log n \leq \log n - \omega(1)$, this instance of parameters is in the converse region given by (47). However, this instance is out of the converse region (45), because

$$\begin{aligned} -\frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) &= -\frac{m}{2}\log(1 - 2(\sqrt{q_{11}q_{00}} - \sqrt{q_{01}q_{10}})^2) \\ &= -\frac{m}{2}\log(1 - 2q_{00}q_{11}) \\ &\geq \frac{m}{2}2q_{00}q_{11} \\ &= (\log n - \log\log n)(1 - o(1)) \\ &> (1 - \Omega(1))\log n. \end{aligned}$$

To make the comparison fair, we comment that there also exists certain converse region given by (45), which is not covered by (47).

## V. Proof of Converse Statement

In this section, we give a detailed proof for Theorem 3. Let $(G_1, G_2)$ be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

In this proof, we will focus on the intersection graph of $G_1$ and $G_2$, denoted as $G_1 \wedge G_2$, which is the graph on the vertex set $\mathcal{V} = \mathcal{V}_{\rm u} \cup \mathcal{V}_{\rm a}$ whose edge set is the intersection of the edge sets of $G_1$ and $G_2$. We say a permutation $\pi$ on the vertex set $\mathcal{V}_{\rm u}$ is an automorphism of $G_1 \wedge G_2$ if a user-user 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 $G_1 \wedge G_2$, and a user-attribute pair $(i, a)$ is in the edge set of $G_1 \wedge G_2$ if and only if $(\pi(i), a)$ is in the edge set of $G_1 \wedge G_2$ i.e., if $\pi$ is *edge-preserving*. Note that an identity permutation is always an automorphism. Let $\mathrm{Aut}(G_1 \wedge G_2)$ denote the set of automorphisms of $G_1 \wedge G_2$. By Lemma 2 below, exact alignment cannot be achieved w.h.p. if $\mathrm{Aut}(G_1 \wedge G_2)$ contains permutations other than the identity permutation. This allows us to establish conditions for not achieving exact alignment w.h.p. by analyzing automorphisms of $G_1 \wedge G_2$.

*Lemma 2:* 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 $\frac{1}{|\mathrm{Aut}(G_1 \wedge G_2)|}$.

The proof of Lemma 2 is deferred to Appendix C.

In the proof of Theorem 3, we will further 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 $|\mathrm{Aut}(G_1 \wedge G_2) \setminus \{\text{identity permutation}\}| \geq |\{\text{indistinguishable vertex pairs}\}|$. In the lemma below, we establish the sharp threshold for the existence of indistinguishable vertex pair.

*Lemma 3 (Sharp Threshold for Indistinguishable Pair):* Let $(G_1, G_2)$ be an attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$ and let $G = G_1 \wedge G_2$. If

$$-n\log(1-2p_{11}+2p_{11}^2) - m\log(1-2q_{11}+2q_{11}^2) - 2\log n \to -\infty, \quad (48)$$

then with high probability, there exists at least one pair of indistinguishable vertices.

Conversely, if

$$-n\log(1-2p_{11}+2p_{11}^2) - m\log(1-2q_{11}+2q_{11}^2) - 2\log n \to \infty, \quad (49)$$

then with high probability, there exists no indistinguishable vertex pairs.

The proof of Lemma 3 is deferred to the end of this section. We complete the proof of Theorem 3 based on Lemma 3 in the following.

*Theorem 10 (Converse):* Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If

$$-n\log(1-2p_{11}+2p_{11}^2) - m\log(1-2q_{11}+2q_{11}^2) - 2\log n \to -\infty, \quad (11)$$

then for any estimator, the probability of error is bounded away from zero.

*Remark 11:* A novelty in the converse proof is that we show the existence of *indistinguishable user pairs* under the attributed Erdős–Rényi graph pair model, while the converse proof in [9] shows the existence of *isolated vertices* under the Erdős–Rényi graph pair model. The reason why we take a different approach from [9] is because (11) cannot guarantee the existence of isolated user vertices with high probability. To see this, let us consider the example of $p_{11} = \frac{0.5\log n}{n}$, $m = \log n$ and $q_{11} = 0.49$. In this case, (11) is satisfied because $-n\log(1-2p_{11}+2p_{11}^2) - m\log(1-2q_{11}+2q_{11}^2) \leq 2np_{11} + 2mq_{11} = 1.98\log n$, where the inequality follows because $-\log(1-2x+2x^2) \leq 2x$ for any $x \geq 0$. It is not hard to see that the expected number of isolated vertices can be calculated as $n(1-p_{11})^{n-1}(1-q_{11})^m$. It follows that

$$\begin{aligned}
&\mathsf{P}(\exists \text{ an isolated vertex }) \\
&\leq \mathsf{E}[\#\text{ isolated vertices}] \\
&= n(1-p_{11})^{n-1}(1-q_{11})^m \\
&\leq \exp(\log n - (n-1)p_{11} + \log 0.51 \times \log n) \\
&= \exp(\log n - 0.5\log n + \log 0.51 \times \log n) \\
&\approx \exp(-0.17\log n) \to 0,
\end{aligned}$$

i.e., the expected number of isolated vertices goes to zero. So with high probability, there exist no isolated vertices in the graph.

*Proof:* [Proof of Theorem 3] Let $X$ denote the number of indistinguishable user vertex pairs in $G$, i.e.,

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

By Lemma 3, we know that $\mathsf{P}(X = 0) = o(1)$. 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 2, $\mathsf{P}(\pi_{\rm MAP} = \Pi^* | X = x) \leq \frac{1}{x+1}$, which is at most $1/2$ when $x \geq 1$. Therefore,

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

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. □

*Proof:* [Proof of Lemma 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.,

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

We will firstly show that $\mathsf{P}(X = 0) \to 0$ as $n \to \infty$ if the condition (48) in Lemma 3 is satisfied.

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

$$\mathsf{P}(X = 0) \leq \frac{\mathsf{Var}(X)}{\mathsf{E}[X]^2} = \frac{\mathsf{E}[X^2] - \mathsf{E}[X]^2}{\mathsf{E}[X]^2}. \quad (50)$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

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

$$\mathsf{E}[X] = \sum_{i<j} \mathsf{P}(i \equiv j) = \binom{n}{2} \mathsf{P}(i \equiv j). \tag{51}$$

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

$$\begin{aligned}
\mathsf{E}[X^2] &= \mathsf{E}\left[\sum_{i<j} \mathbb{1}\{i \equiv j\} \cdot \sum_{k<l} \mathbb{1}\{k \equiv l\}\right] \\
&= \mathsf{E}\left[\sum_{i<j} \mathbb{1}\{i \equiv j\} + \sum_{\substack{i,j,k,l:\ i<j,k<l \\ i,j,k,l \text{ are distinct}}} \mathbb{1}\{i \equiv j\} \cdot \mathbb{1}\{k \equiv l\} \right. \\
&\qquad \left. + \sum_{\substack{i,j,k,l:\ i<j,k<l \\ \{i,j\} \text{ and } \{k,l\} \text{ share one element}}} \mathbb{1}\{i \equiv j \equiv k \equiv l\}\right] \\
&= \binom{n}{2} \mathsf{P}(i \equiv j) + \binom{n}{4}\binom{4}{2} \mathsf{P}(i \equiv j \text{ and } k \equiv l) \\
&\quad + 6\binom{n}{3} \mathsf{P}(i \equiv j \equiv k),
\end{aligned} \tag{52}$$

where $i, j, k, l$ are distinct in (52). With (51) and (52), the upper bound given by Chebyshev's inequality in (50) can be written as

$$\begin{aligned}
\frac{\mathsf{Var}(X)}{\mathsf{E}[X]^2} &= \frac{2}{n(n-1)\mathsf{P}(i \equiv j)} + \frac{4(n-2)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \equiv k)}{\mathsf{P}(i \equiv j)^2} \\
&\quad + \frac{(n-2)(n-3)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2} - 1.
\end{aligned} \tag{53}$$

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

$$\begin{aligned}
&\mathsf{P}(i \equiv j) \\
&= \mathsf{P}(A_1)\mathsf{P}(A_2) \\
&= \left(p_{11}^2 + (1-p_{11})^2\right)^{n-2} \left(q_{11}^2 + (1-q_{11})^2\right)^{m} \\
&= \left(1 - 2p_{11} + 2p_{11}^2\right)^{n-2} \left(1 - 2q_{11} + 2q_{11}^2\right)^{m}.
\end{aligned} \tag{54, 55}$$

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

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

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 (53).

We firstly focus on the fraction $\frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2}$, and show that it converges to zero. Note that condition (48) implies that $-\log(p_{11}^2 + (1-p_{11})^2) = o(1)$. This further implies that $p_{11} = o(1)$ or $1 - p_{11} = o(1)$. As a result, we have $p_{11}^2 + (1-p_{11})^2 = 1 - o(1)$ and $\mathsf{P}(C_0) = 1 - o(1)$. It follows that $\frac{\mathsf{P}(i \equiv j \text{ and } k \equiv l)}{\mathsf{P}(i \equiv j)^2} = \frac{\mathsf{P}(C_0)}{(p_{11}^2 + (1-p_{11})^2)^4} \to 1$. 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 (53). For the first term $\frac{2}{n(n-1)\mathsf{P}(i \equiv j)}$, plugging in the expression in (55) gives

$$\begin{aligned}
&-\log \frac{2}{n(n-1)\mathsf{P}(i \equiv j)} \\
&= 2\log n + (n-2)\log\left(1 - 2p_{11} + 2p_{11}^2\right) \\
&\quad + m\log\left(1 - 2q_{11} + 2q_{11}^2\right) + O(1) \\
&= \omega(1).
\end{aligned} \tag{56}$$

Here (56) follows from condition (48). Therefore, the first term in (53) $\frac{2}{n(n-1)\mathsf{P}(i \equiv j)} \to 0$ 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 (53), we have

$$\begin{aligned}
&-\log\left(\frac{4(n-2)}{n(n-1)} \frac{\mathsf{P}(i \equiv j \equiv k)}{\mathsf{P}(i \equiv j)^2}\right) \\
&= \log n - (n-2)\log\left(\frac{1 - 3p_{11} + 3p_{11}^2}{(1 - 2p_{11} + 2p_{11}^2)^2}\right) \\
&\quad - m\log\left(\frac{1 - 3q_{11} + 3q_{11}^2}{(1 - 2q_{11} + 2q_{11}^2)^2}\right) + O(1) \\
&\geq \log n + \frac{1}{2} n \log(1 - 2p_{11} + 2p_{11}^2)
\end{aligned}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

$$+\frac{1}{2}m\log(1-2q_{11}+2q_{11}^2)+O(1) \tag{57}$$
$$=\omega(1). \tag{58}$$

Here (57) follows from the inequality $-\log\left(\frac{1-3x+3x^2}{(1-2x+2x^2)^2}\right)\geq \frac{1}{2}\log(1-2x+2x^2)$ for any $x\in[0,1]$. Equation (58) follows from the condition (48). Hence, the second term in (53) also converges to 0 as $n\to\infty$, which completes the proof for $\mathsf{P}(X=0)\to 0$ as $n\to\infty$.

Next, we show that $\mathsf{P}(X\geq 1)=o(1)$ under condition (49). From (51) and (55), we know that

$$\mathsf{E}[X]=\binom{n}{2}\left(1-2p_{11}+2p_{11}^2\right)^{n-2}\left(1-2q_{11}+2q_{11}^2\right)^{m}.$$

By Markov's inequality, we have

$$\begin{aligned}\mathsf{P}(X\geq 1)&\leq \mathsf{E}[X]\\&\leq \exp(2\log n+(n-2)\log(1-2p_{11}+2p_{11}^2)\\&\quad+m(1-2q_{11}+2q_{11}^2))\\&=\exp(-\omega(1))=o(1),\end{aligned} \tag{59}$$

where (59) follows by condition (49). □

## VI. Proof of the General Achievability

In this section, we establish the general achievability result in Theorem 1. In this proof, we first simplify the optimal estimator for exact alignment, the MAP estimator, to a minimum weighted distance estimator in Lemma 4. We then analyze the probability analyze the error probability that a wrong permutation $\pi$ has a lower weighted distance than the true underlying permutation in Lemma 5. The main idea in the proof of Lemma 5 is first bounding the error probability with the probability generating function of the difference of weighted distance. The generating function is then further bounded by applying a cycle decomposition to the edge permutation induced by $\pi$. With these two key lemmas, the proof of Theorem 1 is then completed by a straightforward union bound argument. In the following, we state the two key lemmas and prove Theorem 1 based on them.

To state Lemma 4, we first introduce some basic notation for graph statistics needed in stating the MAP estimator. For any attributed graph $g$ on the vertex set $\mathcal{V}_{\mathrm{u}}\cup\mathcal{V}_{\mathrm{a}}$ and any permutation $\pi$ over the user vertex set $\mathcal{V}_{\mathrm{u}}$, we use $\pi(g)$ to denoted the graph given by applying $\pi$ 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

$$\Delta^{\mathrm{u}}(g_1,g_2)=\sum_{(i,j)\in\mathcal{E}_{\mathrm{u}}}\mathbb{1}\{g_1((i,j))\neq g_2((i,j))\}; \tag{60}$$

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

$$\Delta^{\mathrm{a}}(g_1,g_2)=\sum_{(i,v)\in\mathcal{E}_{\mathrm{a}}}\mathbb{1}\{g_1((i,v))\neq g_2((i,v))\}. \tag{61}$$

*Lemma 4:* 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 $\Pi^*$ 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

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

In the following lemma, we upper bound the probability that a permutation has a lower weighted distance than the identity permutation.

*Lemma 5:* Let $(G_1,G_2)$ be an attributed Erdős–Rényi pair $\mathcal{G}(n,\boldsymbol{p};m,\boldsymbol{q})$. For any permutation $\pi$, 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 $\pi$ has $n-\tilde{n}$ fixed points, we have

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

We defer the proof of Lemma 4 to Appendices A, and defer the proof of Lemma 5 to the end of this section. With these two lemmas, we are now ready to prove Theorem 1. *Proof:* [Proof of Theorem 1] Given the observable pair $(G_1,G_2')$, the error probability of MAP estimator can be upper-bounded as

$$\begin{aligned}&\mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1,G_2')\neq\Pi^*)\\&=\sum_{\pi^*\in\mathcal{S}_n}\mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1,G_2')\neq\pi^*|\Pi^*=\pi^*)\mathsf{P}(\Pi^*=\pi^*)\\&=\frac{1}{|\mathcal{S}_n|}\sum_{\pi^*\in\mathcal{S}_n}\mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1,G_2')\neq\pi^*|\Pi^*=\pi^*) &(62)\\&=\mathsf{P}\left(\hat{\pi}_{\mathrm{MAP}}(G_1,G_2)\neq\pi_{\mathrm{id}}|\Pi^*=\pi_{\mathrm{id}}\right) &(63)\\&=\mathsf{P}(\hat{\pi}_{\mathrm{MAP}}(G_1,G_2)\neq\pi_{\mathrm{id}}) &(64)\\&\leq\mathsf{P}(\exists\pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\},\delta_\pi(G_1,G_2)\leq 0), &(65)\end{aligned}$$

where $\pi_{\mathrm{id}}$ denotes the identity permutation, and

$$\begin{aligned}\delta_\pi(G_1,G_2)\triangleq\ &w_1(\Delta^{\mathrm{u}}(G_1,\pi(G_2))-\Delta^{\mathrm{u}}(G_1,G_2))\\&+w_2(\Delta^{\mathrm{a}}(G_1,\pi(G_2))-\Delta^{\mathrm{a}}(G_1,G_2)).\end{aligned} \tag{66}$$

Here (62) follows from the fact that $\Pi^*$ is uniformly drawn from $\mathcal{S}_n$, which implies $\mathsf{P}(\Pi^*=\pi^*)=1/|\mathcal{S}_n|$ for all $\pi^*$; (63) is due to the symmetry among user vertices in $G_1$ and $G_2$; (64) is due to the independence between $\Pi^*$ and $(G_1,G_2)$; (65) is true because by Lemma 4, $\pi_{\mathrm{MAP}}(G_1,G_2)$ minimizes the weighted distance, and $\pi_{\mathrm{MAP}}\neq\pi_{\mathrm{id}}$ only if there exists a permutation $\pi$ such that $\pi\neq\pi_{\mathrm{id}}$ and $\delta_\pi(G_1,G_2)\leq 0$.

Now to prove that (6) implies that the error probability in (65) converges to 0 as $n\to\infty$, we further upper-bound the error probability as follows

$$\begin{aligned}&\mathsf{P}\left(\exists\pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\},\delta_\pi(G_1,G_2)\leq 0\right)\\&\leq\sum_{\pi\in\mathcal{S}_n\setminus\{\pi_{\mathrm{id}}\}}\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0)\end{aligned} \tag{67}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

$$= \sum_{\tilde{n}=2}^{n} \sum_{\pi \in \mathcal{S}_{n,\tilde{n}}} \mathsf{P}(\delta_\pi(G_1, G_2) \le 0) \tag{68}$$
$$\le \sum_{\tilde{n}=2}^{n} |\mathcal{S}_{n,\tilde{n}}| \max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \le 0)\}$$
$$\le \sum_{\tilde{n}=2}^{n} n^{\tilde{n}} \max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \le 0)\}.$$

Here (67) follows from directly applying the union bound. In (68), 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, the given permutation $\Pi^* = (1)(23)$ has 1 fixed point and $(1)(23) \in \mathcal{S}_{3,2}$. Furthermore, we have $|\mathcal{S}_{n,\tilde{n}}| = \binom{n}{\tilde{n}} (!\tilde{n}) \le n^{\tilde{n}}$, 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.

With the upper bound in Lemma 5, we have

$$\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0\right)$$
$$\le \sum_{\tilde{n}=2}^{n} n^{\tilde{n}} (1-2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}} (1-2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}}$$
$$= \sum_{\tilde{n}=2}^{n} \left( n(1-2\psi_{\mathrm{u}})^{\frac{n-2}{4}} (1-2\psi_{\mathrm{a}})^{\frac{m}{2}} \right)^{\tilde{n}}. \tag{69}$$

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

$$-\log n - \frac{n-2}{4} \log(1-2\psi_{\mathrm{u}}) - \frac{m}{2} \log(1-2\psi_{\mathrm{a}})$$
$$\ge -\log n + \frac{n-2}{2} \psi_u + m\psi_a \tag{70}$$
$$= \omega(1). \tag{71}$$

Here we have $\psi_{\mathrm{u}} = (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \le 1/4$ and $\psi_{\mathrm{a}} = (\sqrt{q_{11}q_{00}} - \sqrt{q_{10}q_{01}})^2 \le 1/4$. Therefore, (70) follows from the inequality $\log(1+x) \le x$ for $x > -1$. Equation (71) follows from condition (6) by noting that $\psi_u$ is no larger than 1. Therefore, the geometry series in (69) converge to 0 as $n \to \infty$. This completes the proof that MAP estimator achieves exact alignment w.h.p. under condition (6). □

*Proof:* [Proof of Lemma 5] To prove the upper bound on $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0)$ in Lemma 5, we will use the method of *generating functions*. We first introduce our construction of a generating function and how it can be used to bound $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0)$. We then present several properties of generating functions (Facts 1, 2, and 3), which will be needed to finish the proof of Lemma 5.

*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 $2 \times 2$ 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 $2 \times 2$ 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 $\pi$, we construct the following generating function:

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) = \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}, \tag{72}$$

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 $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0)$ 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 $\mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z) = \mathsf{E}\left[z^{\delta_\pi(G_1,G_2)}\right]$. 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

$$\mathsf{P}(\delta_\pi(G_1, G_2) \le 0) = \sum_{d \le 0} [z^d] \mathcal{A}(\boldsymbol{p}, \boldsymbol{q}, z). \tag{73}$$

*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 $\pi$ 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

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

from a cycle. We refer to each of these subsets as an *orbit*. For the example given in Figure 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 $k$th 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 $\delta_\pi$, $\boldsymbol{\mu}$ 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 $\delta_\pi$ for any $g,h\in\{0,1\}^{\mathcal{E}'}$ as

$$\begin{aligned}\delta_\pi(g,h)=&w_1\sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{u}}}\left(\mathbb{1}\{g(e)\neq h(\pi^{\mathcal{E}}(e))\}-\mathbb{1}\{g(e)\neq h(e)\}\right)\\&+w_2\sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{a}}}\left(\mathbb{1}\{g(e)\neq h(\pi^{\mathcal{E}}(e))\}-\mathbb{1}\{g(e)\neq h(e)\}\right).\end{aligned}\tag{74}$$

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

$$\mu_{ij}=\mu_{ij}(g,h)\triangleq\sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{u}}}\mathbb{1}\{g(e)=i,h(e)=j\},\tag{75}$$

$$\nu_{ij}=\nu_{ij}(g,h)\triangleq\sum_{e\in\mathcal{E}'\cap\mathcal{E}_{\mathrm{a}}}\mathbb{1}\{g(e)=i,h(e)=j\}.\tag{76}$$

We remind the reader that by setting the set of vertex pairs $\mathcal{E}'$ to be $\mathcal{E}$ these extended definitions on $\delta_\pi$, $\boldsymbol{\mu}$ 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

$$\mathcal{A}_l(\boldsymbol{x},z)\triangleq\sum_{g\in\{0,1\}^{\mathcal{E}_l^{\mathrm{u}}}}\sum_{h\in\{0,1\}^{\mathcal{E}_l^{\mathrm{u}}}} z^{\delta_\pi(g,h)}\boldsymbol{x}^{\boldsymbol{\mu}(g,h)},\tag{77}$$

$$\mathcal{A}_l(\boldsymbol{y},z)\triangleq\sum_{g\in\{0,1\}^{\mathcal{E}_l^{\mathrm{a}}}}\sum_{h\in\{0,1\}^{\mathcal{E}_l^{\mathrm{a}}}} z^{\delta_\pi(g,h)}\boldsymbol{y}^{\boldsymbol{\nu}(g,h)}.\tag{78}$$

*Properties of Generating Functions:*

*Fact 1:* The generating function $\mathcal{A}(\boldsymbol{x},\boldsymbol{y},z)$ of permutation $\pi$ 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 $t_l^{\mathrm{u}}$ is the number of user-user orbits of size $l$, $t_l^{\mathrm{a}}$ 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 D for the proof of Fact 1, and Theorem 4 in [9] for the proof of Fact 2. Combining these two facts, we get

$$\begin{aligned}\mathcal{A}(\boldsymbol{x},\boldsymbol{y},z)\leq&\mathcal{A}_1(\boldsymbol{x},z)^{t_1^{\mathrm{u}}}\mathcal{A}_1(\boldsymbol{y},z)^{t_1^{\mathrm{a}}}\\&\mathcal{A}_2(\boldsymbol{x},z)^{\frac{t^{\mathrm{u}}-t_1^{\mathrm{u}}}{2}}\mathcal{A}_2(\boldsymbol{x},z)^{\frac{nm-t_1^{\mathrm{a}}}{2}}.\end{aligned}\tag{79}$$

Here, in (79), we use $t^{\mathrm{u}}$ 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 (77) and (78)

$$\mathcal{A}_1(\boldsymbol{x},z)=x_{00}+x_{10}+x_{01}+x_{11},\tag{80}$$

$$\mathcal{A}_1(\boldsymbol{y},z)=y_{00}+y_{10}+y_{01}+y_{11},\tag{81}$$

$$\begin{aligned}\mathcal{A}_2(\boldsymbol{x},z)=&(x_{00}+x_{10}+x_{01}+x_{11})^2\\&+2x_{00}x_{11}(z^{2w_1}-1)+2x_{10}x_{01}(z^{-2w_1}-1),\end{aligned}\tag{82}$$

$$\begin{aligned}\mathcal{A}_2(\boldsymbol{y},z)=&(y_{00}+y_{10}+y_{01}+y_{11})^2\\&+2y_{00}y_{11}(z^{2w_2}-1)+2y_{10}y_{01}(z^{-2w_2}-1).\end{aligned}\tag{83}$$

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

$$\Phi(z)\triangleq\mathsf{E}[z^X]=\sum_{i\in\mathcal{X}}\mathsf{P}(X=i)z^i\tag{84}$$

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

$$[z^j]\Phi(z)\leq z^{-j}\Phi(z).\tag{85}$$

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

$$\sum_{\substack{i\leq j\\ i\in\mathcal{X}}}[z^i]\Phi(z)\leq z^{-j}\Phi(z).\tag{86}$$

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

$$\sum_{\substack{i\geq j\\ i\in\mathcal{X}}}[z^i]\Phi(z)\leq z^{-j}\Phi(z).\tag{87}$$

*Proof:* [Proof of Fact 3] We write $p_i\triangleq\mathsf{P}(X=i)$ in this proof. For any $j\in\mathcal{X}$ and $z>0$, we have

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

which establishes (85).

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

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

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

which establishes (86).

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

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

which establishes (87). □

With the three facts of generating functions stated above, we are now ready to finish the proof of Lemma 5. For any $\pi \in \mathcal{S}_{n,\tilde{n}}$ and any $z_1 \in (0,1)$, we have

$$\begin{aligned}
&\mathsf{P}\left(\delta_\pi(G_1,G_2)\le 0\right)\\
&= \sum_{d\le 0}[z^d]\mathcal{A}(\boldsymbol{p},\boldsymbol{q},z)\\
&\le \mathcal{A}(\boldsymbol{p},\boldsymbol{q},z_1) \qquad (88)\\
&\le \mathcal{A}_1(\boldsymbol{p},z)^{t_1^{\mathrm{u}}}\mathcal{A}_1(\boldsymbol{q},z)^{t_1^{\mathrm{a}}}\\
&\quad \mathcal{A}_2(\boldsymbol{p},z)^{\frac{t^{\mathrm{u}}-t_1^{\mathrm{u}}}{2}}\mathcal{A}_2(\boldsymbol{q},z)^{\frac{nm-t_1^{\mathrm{a}}}{2}} \qquad (89)\\
&\le \mathcal{A}_2(\boldsymbol{p},z)^{\frac{t^{\mathrm{u}}-t_1^{\mathrm{u}}}{2}}\mathcal{A}_2(\boldsymbol{q},z)^{\frac{nm-t_1^{\mathrm{a}}}{2}}. \qquad (90)
\end{aligned}$$

In (88), we set $z \in (0,1)$, and this upper bound follows from Fact 3. (89) follows from the decomposition on $\mathcal{A}(\boldsymbol{p},\boldsymbol{q},z)$ stated in Fact 1. Equation (90) follows since $\mathcal{A}_1(\boldsymbol{p},z) = \mathcal{A}_1(\boldsymbol{q},z) = 1$ according to their expression in (80) and (81). To obtain a tight bound, we then search for $z \in (0,1)$ that achieves the minimum of (90). Following the definition of $\mathcal{A}_2(\boldsymbol{p},z)$ in (82) and using the inequality $a/x + bx \ge 2\sqrt{ab}$, we have

$$\begin{aligned}
\mathcal{A}_2(\boldsymbol{p},z) &= 1 + 2p_{00}p_{11}(z^{2w_1}-1) + 2p_{10}p_{01}(z^{-2w_1}-1)\\
&\ge 1 - 2p_{00}p_{11} - 2p_{10}p_{01} + 4\sqrt{p_{00}p_{11}p_{10}p_{01}}\\
&= 1 - 2(\sqrt{p_{00}p_{11}} - \sqrt{p_{10}p_{01}})^2 \triangleq 1 - 2\psi_{\mathrm{u}}. \qquad (91)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{A}_2(\boldsymbol{q},z) &= 1 + 2q_{00}q_{11}(z^{2w_2}-1) + 2q_{10}q_{01}(z^{-2w_2}-1)\\
&\ge 1 - 2q_{00}q_{11} - 2q_{10}q_{01} + 4\sqrt{q_{00}q_{11}q_{10}q_{01}}\\
&= 1 - 2(\sqrt{q_{00}q_{11}} - \sqrt{q_{10}q_{01}})^2 \triangleq 1 - 2\psi_{\mathrm{a}}. \qquad (92)
\end{aligned}$$

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 (90) and we have

$$\begin{aligned}
&\mathsf{P}\left(\delta_\pi(G_1,G_2)\le 0\right)\\
&\le (1-2\psi_{\mathrm{u}})^{\frac{t^{\mathrm{u}}-t_1^{\mathrm{u}}}{2}}(1-2\psi_{\mathrm{a}})^{\frac{mn-t_1^{\mathrm{a}}}{2}}\\
&\le (1-2\psi_{\mathrm{u}})^{\frac{\tilde{n}(2n-\tilde{n}-2)}{4}}(1-2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}} \qquad (93)\\
&\le (1-2\psi_{\mathrm{u}})^{\frac{\tilde{n}(n-2)}{4}}(1-2\psi_{\mathrm{a}})^{\frac{\tilde{n}m}{2}}. \qquad (94)
\end{aligned}$$

In (93), we use the following relations between the number of fixed vertex pairs $t_1^{\mathrm{u}}$, $t_1^{\mathrm{a}}$ and number of fixed vertices $\tilde{n}$

$$\begin{aligned}
&\binom{n-\tilde{n}}{2} \le t_1^{\mathrm{u}} \le \binom{n-\tilde{n}}{2} + \frac{\tilde{n}}{2}, \qquad (95)\\
&t_1^{\mathrm{a}} = (n-\tilde{n})m.
\end{aligned}$$

In the given upper bound of $t_1^{\mathrm{u}}$, $\binom{n-\tilde{n}}{2}$ corresponds to the number of user-user vertex pairs whose two vertices are both fixed under $\pi$, and $\frac{\tilde{n}}{2}$ is the upper bound of user-user vertex pairs whose two vertices are swapped under $\pi$. In (94), we use the fact that $\tilde{n} \le n$. □

## VII. Proof of Achievability in the Sparse Regime

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 Section VI. The reason for applying a different proof technique is that, in this sparse regime, the union bound we applied in Section VI on $\mathsf{P}\left(\exists\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1,G_2)\le 0\right)$ 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 $\{\delta_\pi(G_1,G_2) \le 0\}$ for all $\pi \in \mathcal{S}_n$. Moreover, other events where $R$ is small are also in the intersection of $\{\delta_\pi(G_1,G_2) \le 0\}$ 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

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

We then apply the union bound on the conditional probability $\mathsf{P}\left(\exists\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1,G_2)\le 0 | R=r\right)$. 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 $\mathsf{P}\left(\exists\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1,G_2)\le 0 | R=r\right)$ 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

$$\mathsf{P}\left(\exists\pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1,G_2)\le 0 | R=r\right)$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

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

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 $\mathsf{P}\left(\delta_\pi(G_1, G_2) \le 0 | R = 0\right) = 1$ 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 $\mathsf{P}\left(\delta_\pi(G_1, G_2) \le 0 | R = r\right)$ 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 [9] and is extended to the attributed Erdős–Rényi pair model. We restate the theorem to prove as follows.

*Theorem 11 (Achievability in Sparse Region):* Consider the attributed Erdős–Rényi pair $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$. If

$$p_{11} = O\left(\tfrac{\log n}{n}\right), \tag{7}$$

$$p_{10} + p_{01} = O\left(\tfrac{1}{\log n}\right), \tag{8}$$

$$\frac{p_{10}p_{01}}{p_{11}p_{00}} = O\left(\tfrac{1}{(\log n)^3}\right), \tag{9}$$

$$np_{11} + m\psi_{\mathrm{a}} - \log n = \omega(1), \tag{10}$$

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

*Proof:* [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 (6) for exact alignment in Theorem 1

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

is satisfied when condition (10)

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

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

$$\begin{aligned}
&\mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0\right) \\
&= \sum_{r=0}^{t^{\mathrm{u}}} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) \\
&\le \sum_{r \le (1+\epsilon)\mathsf{E}[R]} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) \\
&\qquad + \mathsf{P}\left(|R - \mathsf{E}[R]| > \epsilon \mathsf{E}[R]\right) \\
&= \sum_{r \le (1+\epsilon)\mathsf{E}[R]} \mathsf{P}\left(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 | R = r\right) \mathsf{P}(R = r) \\
&\qquad + o(1) \qquad (96) \\
&\overset{\text{Lemma 6}}{\le} \sum_{r \le (1+\epsilon)\mathsf{E}[R]} 3n^2 z_6^2 \mathsf{P}(R = r) + o(1) \qquad (97) \\
&= \sum_{r \le (1+\epsilon)\mathsf{E}[R]} 3n^2 z_6^2 \binom{t^{\mathrm{u}}}{r} p_{11}^r (1 - p_{11})^{t^{\mathrm{u}} - r} + o(1) \\
&= 3n^2 (1 - 2\psi_{\mathrm{a}})^m \sum_{r=0}^{t^{\mathrm{u}}} \binom{t^{\mathrm{u}}}{r} p_{11}^r e^{-\frac{4r}{n}} (1 - p_{11})^{t^{\mathrm{u}} - r} + o(1) \\
&= 3n^2 (1 - 2\psi_{\mathrm{a}})^m \left(p_{11} e^{-\frac{4}{n}} + 1 - p_{11}\right)^{t^{\mathrm{u}}} + o(1) \qquad (98) \\
&\le 3n^2 (1 - 2\psi_{\mathrm{a}})^m \left(1 - \tfrac{4}{n} p_{11}\right)^{t^{\mathrm{u}}} + o(1). \qquad (99)
\end{aligned}$$

Here (96) follows from the Chebyshev's inequality above. In (97), $z_6 = \exp\{-\frac{2r}{n} + \frac{m}{2}\log(1 - 2\psi_{\mathrm{a}}) + O(1)\}$. This step will be justified by Lemma 6, which is the major technical step in establishing the error bound. To apply Lemma 6, we need the conditions (7) (8) (9) and $r = O(\mathsf{E}[R]) = O(n \log n)$ to hold and we will explain the reason in the proof of Lemma 6. Equation (98) follows from the binomial formula and (99) follows from the inequality $e^x - 1 \le x$. Taking the negative logarithm of the first term in (99), we have

$$\begin{aligned}
&-\log\left(3n^2(1 - 2\psi_{\mathrm{a}})^m \left(1 - \tfrac{4}{n}p_{11}\right)^{t^{\mathrm{u}}}\right) \\
&= -2\log n - m\log(1 - 2\psi_{\mathrm{a}}) - t^{\mathrm{u}} \log\left(1 - \tfrac{4p_{11}}{n}\right) + O(1) \\
&\ge -2\log n + 2m\psi_{\mathrm{a}} + t^{\mathrm{u}} \frac{4p_{11}}{n} + O(1) \qquad (100) \\
&= -2\log n + 2m\psi_{\mathrm{a}} + 2np_{11} + O(1) \qquad (101) \\
&= \omega(1). \qquad (102)
\end{aligned}$$

Here, we have (100) follows from the inequality $\log(1 + x) \le x$ for $x > -1$. We get equation (101) by plugging in $t^{\mathrm{u}} = \binom{n}{2}$. Equation (102) follows from the assumption (10) in Theorem 2. Therefore, we have (99) converges to 0 and so does the error probability. □

*Lemma 6:* 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 (7) (8) (9), and $r = O(n \log n)$, then

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

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 $\mathcal{E}_{\mathrm{m}} = \{e \in \mathcal{E} : \pi^{\mathcal{E}}(e) \ne e\}$. 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 $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0 \mid \tilde{R} = \tilde{r})$. The reason for conditioning on $\tilde{R}$ first is that the corresponding generating function only involves cycles of length $l \ge 2$ and its upper bound is easier to derive compared with the probability conditioned on $R$. In Step 2, we upper bound $\mathsf{P}(\delta_\pi(G_1, G_2) \le 0 \mid R = r)$ using result from Step 1 and properties of the Hypergeometric distribution. In Step 3, we upper bound $\mathsf{P}(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \le 0 \mid R = r)$ using the truncated union bound.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

*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 $z_3 = (1-2\psi_{\mathrm{a}})^{\frac{m}{2}}$,

$$\mathsf{P}\left(\delta_\pi(G_1,G_2) \le 0 \mid \tilde{R} = \tilde{r}\right) \le z_3^{\tilde{n}} z_4^{\tilde{r}} z_5^{\tilde{n}} \tag{103}$$

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

$$\tilde{\mathcal{A}}(\boldsymbol{x},\boldsymbol{y},z) = \sum_{g\in\{0,1\}^{\mathcal{E}_{\mathrm{m}}}} \sum_{h\in\{0,1\}^{\mathcal{E}_{\mathrm{m}}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)}. \tag{104}$$

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 (74), (75) and (76). We have

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

For the $2\times 2$ 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

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

where $t_l^{\mathrm{u}}$ is the number of user-user orbits of size $l$ and $t_l^{\mathrm{a}}$ 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 (104). 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 an 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

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

Thus, we can write

$$\begin{aligned}&\mathsf{P}\left(\delta_\pi(G_1,G_2)\le 0, \tilde{R}=\tilde{r}\right) \\ &= \sum_{d\le 0} [z^d x_{11}^{\tilde{r}}] \tilde{\mathcal{A}}(\boldsymbol{x_{11}},\boldsymbol{q},z) \\ &= \sum_{d\le 0} [z^d x_{11}^{\tilde{r}}] \prod_{l\ge 2} \mathcal{A}_l(\boldsymbol{x_{11}},z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{q},z)^{t_l^{\mathrm{a}}} \\ &\le (x_{11})^{-\tilde{r}} \sum_{d\le 0} [z^d] \prod_{l\ge 2} \mathcal{A}_l(\boldsymbol{x_{11}},z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{q},z)^{t_l^{\mathrm{a}}} \qquad (105) \\ &\le (x_{11})^{-\tilde{r}} \prod_{l\ge 2} \mathcal{A}_l(\boldsymbol{x_{11}},z)^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{q},z)^{t_l^{\mathrm{a}}} \qquad (106) \\ &\le (x_{11})^{-\tilde{r}} \mathcal{A}_2(\boldsymbol{x_{11}},z)^{\frac{\tilde{t}^{\mathrm{u}}}{2}} \mathcal{A}_2(\boldsymbol{q},z)^{\frac{m\tilde{n}}{2}}. \qquad (107)\end{aligned}$$

In (105), we set $x_{11}>0$ and the inequality follows from (85) in Fact 3. In (106), we set $z\in(0,1)$ and this inequality follows from (86) Fact 3. Inequality in (107) follows from Fact 2, where

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

is the number of moving user-user pairs and

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

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

$$\begin{aligned}\mathsf{P}(\tilde{R}=\tilde{r}) &= \binom{\tilde{t}^{\mathrm{u}}}{\tilde{r}} p_{11}^{\tilde{r}} (1-p_{11})^{\tilde{t}^{\mathrm{u}}-\tilde{r}} \\ &\ge \left(\frac{\tilde{t}^{\mathrm{u}} p_{11}}{\tilde{r}(1-p_{11})}\right)^{\tilde{r}} (1-p_{11})^{\tilde{t}^{\mathrm{u}}}, \qquad (108)\end{aligned}$$

where equation (108) follows since $\binom{n}{k} \ge (n/k)^k$ for any nonnegative integers $k\le n$.

Now we combine the bounds in (107) and (108) to upper bound $\mathsf{P}\left(\delta_\pi(G_1,G_2)\le 0 \mid \tilde{R}=\tilde{r}\right)$. Define $p'_{ij} \triangleq \frac{p_{ij}}{1-p_{11}}$ for $i,j\in\{0,1\}$. We have

$$\begin{aligned}&\mathsf{P}\left(\delta_\pi(G_1,G_2)\le 0 \mid \tilde{R}=\tilde{r}\right) = \frac{\mathsf{P}(\delta_\pi(G_1,G_2)\le 0, \tilde{R}=\tilde{r})}{\mathsf{P}(\tilde{R}=\tilde{r})} \\ &\le \mathcal{A}_2(\boldsymbol{q},z)^{\frac{m\tilde{n}}{2}} \left(\frac{\tilde{r}}{x_{11}p'_{11}\tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}} \left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}},z)}{(1-p_{11})^2}\right)^{\tilde{t}^{\mathrm{u}}/2}. \qquad (109)\end{aligned}$$

For the first term, similar to what we did in (92), 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 (83), we have

$$\begin{aligned}&\mathcal{A}_2(\boldsymbol{q},z)^{\frac{m\tilde{n}}{2}} \\ &= \left(1+2q_{00}q_{11}(z^{2w_2}-1) + 2q_{10}q_{01}(z^{-2w_2}+1)\right)^{\frac{m\tilde{n}}{2}} \qquad (110) \\ &= \left(1-2q_{00}q_{11}-2q_{10}q_{01}+4\sqrt{q_{00}q_{11}q_{10}q_{01}}\right)^{\frac{m\tilde{n}}{2}} \qquad (111) \\ &= \left(1-2(\sqrt{q_{11}q_{00}}-\sqrt{q_{10}q_{01}})^2\right)^{\frac{m\tilde{n}}{2}} \\ &= (1-2\psi_{\mathrm{a}})^{\frac{m\tilde{n}}{2}} \\ &\triangleq z_3^{\tilde{n}}. \qquad (112)\end{aligned}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

where (110) follows since $q_{00}+q_{01}+q_{10}+q_{11}=1$ and (111) 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 (109), we set

$$x_{11}=\frac{\tilde{r}\log n+p_{11}\tilde{t}^{\mathrm{u}}}{p'_{11}\tilde{t}^{\mathrm{u}}}, \tag{113}$$

which is positive. Then, we have

$$\left(\frac{\tilde{r}}{x_{11}p'_{11}\tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}}=\left(\frac{\tilde{r}}{\tilde{r}\log n+p_{11}\tilde{t}^{\mathrm{u}}}\right)^{\tilde{r}}\le\left(\frac{1}{\log n}\right)^{\tilde{r}}. \tag{114}$$

For the third term in (109), using equation (82) with $z=e^{-1/4}$, we have

$$\begin{aligned}
&\frac{\mathcal{A}_2(\boldsymbol{x_{11}},z)}{(1-p_{11})^2}\\
&=\frac{(1-p_{11}+x_{11}p_{11})^2}{(1-p_{11})^2}+\frac{2x_{11}p_{11}p_{00}(\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}}-1)}{(1-p_{11})^2}\\
&\quad+\frac{2p_{10}p_{01}(\sqrt{\frac{p_{00}p_{11}}{p_{10}p_{01}}}-1)}{(1-p_{11})^2}\\
&=(1+p'_{11}x_{11})^2-2x_{11}p'_{11}p'_{00}-2p'_{10}p'_{01}\\
&\quad+2(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}\\
&\le 1+(p'_{11}x_{11})^2+2p'_{11}x_{11}(p'_{10}+p'_{01})\\
&\quad+2(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}},
\end{aligned}$$

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

$$\begin{aligned}
&\frac{\tilde{t}^{\mathrm{u}}}{2}\log\left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}},z)}{(1-p_{11})^2}\right) &(115)\\
&\le\frac{1}{2}\tilde{t}^{\mathrm{u}}(p'_{11}x_{11})^2+\tilde{t}^{\mathrm{u}}p'_{11}x_{11}(p'_{10}+p'_{01})\\
&\quad+\tilde{t}^{\mathrm{u}}(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}, &(116)
\end{aligned}$$

where (116) follows from the inequality $\log(1+x)\le x$. Let us now bound the three terms in (116).

- For the first term, we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}(p'_{11}x_{11})^2\\
&=\tilde{t}^{\mathrm{u}}\left(\frac{\tilde{r}\log n}{\tilde{t}^{\mathrm{u}}}+p_{11}\right)^2\\
&=\frac{\tilde{r}^2(\log n)^2}{\tilde{t}^{\mathrm{u}}}+2\tilde{r}(\log n)p_{11}+\tilde{t}^{\mathrm{u}}p_{11}^2\\
&=O\left(\frac{\tilde{r}(\log n)^3}{n}\right)+2\tilde{r}(\log n)p_{11}+\tilde{t}^{\mathrm{u}}p_{11}^2 &(117)\\
&=O\left(\frac{\tilde{r}(\log n)^3}{n}+\frac{\tilde{r}(\log n)^2}{n}+\frac{\tilde{n}(\log n)^2}{n}\right) &(118)\\
&=o(\tilde{r}+\tilde{n}),
\end{aligned}$$

  where (117) follows from the assumption $\tilde{r}=O(\frac{\tilde{t}^{\mathrm{u}}\log n}{n})$ in (103) and (118) follows since $p_{11}=O(\frac{\log n}{n})$ and $\tilde{t}^{\mathrm{u}}\le\tilde{n}n$.

- For the second term in (116), we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}p'_{11}x_{11}(p'_{10}+p'_{01})\\
&=(\tilde{r}\log n+\tilde{t}^{\mathrm{u}}p_{11})(p'_{10}+p'_{01})\\
&\le\frac{\tilde{r}(p_{10}+p_{01})\log n+\tilde{n}np_{11}(p_{10}+p_{01})}{1-p_{11}} &(119)\\
&=O(\tilde{r}+\tilde{n}), &(120)
\end{aligned}$$

  where (119) follows from $\tilde{t}^{\mathrm{u}}\le\tilde{n}n$ and (120) 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 (116), we have

$$\begin{aligned}
&\tilde{t}^{\mathrm{u}}(x_{11}+1)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}\\
&=\tilde{t}^{\mathrm{u}}\left(\frac{\tilde{r}\log n+p_{11}\tilde{t}^{\mathrm{u}}}{p'_{11}\tilde{t}^{\mathrm{u}}}+1\right)\sqrt{p'_{11}p'_{00}p'_{10}p'_{01}}\\
&=(\tilde{r}\log n+p_{11}\tilde{t}^{\mathrm{u}}+p'_{11}\tilde{t}^{\mathrm{u}})p'_{00}\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}}\\
&\le(\tilde{r}\log n+p_{11}n\tilde{n}+p'_{11}n\tilde{n})p'_{00}\sqrt{\frac{p_{10}p_{01}}{p_{11}p_{00}}} &(121)\\
&=o(\tilde{r}+\tilde{n}). &(122)
\end{aligned}$$

  Here (121) follows since $\tilde{t}^{\mathrm{u}}\le\tilde{n}n$. (122) 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 (109) is upper bounded as

$$\left(\frac{\mathcal{A}_2(\boldsymbol{x_{11}},z)}{(1-p_{11})^2}\right)^{\tilde{t}^{\mathrm{u}}/2}\le\exp\{O(\tilde{r}+\tilde{n})\}. \tag{123}$$

Finally, combining (112) (114) (123), 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 $r=O(n\log n)$,

$$\mathsf{P}(\delta_\pi(G_1,G_2)\le 0\mid R=r)\le z_6^{\tilde{n}} \tag{124}$$

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 $\Phi_{\mathrm{Hyp}}(z)\le\Phi_{\mathrm{Bin}}(z)$ [27].
- $\Phi_{\mathrm{Bin}}(z)=\left(1+\frac{K}{N}(z-1)\right)^n$.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

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 [9, Lemma IV.5], we have the following upper bound on $\Phi_{\mathrm{Hyp}}(z)$ for any $z \in (0,1)$

$$\Phi_{\mathrm{Hyp}}(z) \leq \exp\left\{\tfrac{r\tilde{n}}{n}(-2+\tfrac{e}{n-1}+2ez)\right\}. \tag{125}$$

Now, we are ready for proving (124). We first write

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0 \mid R=r)\\ &=\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}\leq \tilde{r}^* \mid R=r)\\ &+\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}> \tilde{r}^* \mid R=r).\end{aligned} \tag{126}$$

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 (126), we have

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}\leq \tilde{r}^* \mid R=r)\\ &=\sum_{\tilde{r}\leq\tilde{r}^*}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r)\ \ \mathsf{P}(\delta_\pi(G_1,G_2)\leq 0\mid \tilde{R}=\tilde{r})\end{aligned} \tag{127}$$

$$\leq \sum_{\tilde{r}\leq\tilde{r}^*}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r) z_3^{\tilde{n}} z_4^{\tilde{r}} z_5^{\tilde{n}} \tag{128}$$

$$\begin{aligned}&\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \sum_{\tilde{r}=0}^{n}\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r) z_4^{\tilde{r}}\\ &= z_3^{\tilde{n}} z_5^{\tilde{n}} \Phi_{\mathrm{Hyp}}(z_4)\end{aligned} \tag{129}$$

$$\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{\tfrac{\tilde{n}r}{n}(-2+\tfrac{e}{n-1}+2ez_4)\right\} \tag{130}$$

$$= z_3^{\tilde{n}}(e^{O(1)})^{\tilde{n}} \exp\left\{-\tfrac{2\tilde{n}r}{n}+\tfrac{e\tilde{n}r}{n(n-1)}+O(\tfrac{1}{\log n})\tfrac{\tilde{n}r}{n}\right\} \tag{131}$$

$$\leq z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(-\tfrac{2r}{n}+O(1)\right)\right\} \tag{132}$$

In (127), we use the conditional independence of $R$ and $\delta_\pi(G_1,G_2)$ given $\tilde{R}$, which can be proved as follows

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0|\tilde{R}=\tilde{r},R=r)\\ &=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}=\tilde{r},R=r)}{\mathsf{P}(\tilde{R}=\tilde{r},R=r)}\\ &=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}=\tilde{r},R-\tilde{R}=r-\tilde{r})}{\mathsf{P}(\tilde{R}=\tilde{r},R-\tilde{R}=r-\tilde{r})}\\ &=\frac{\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}=\tilde{r})\mathsf{P}(R-\tilde{R}=r-\tilde{r})}{\mathsf{P}(\tilde{R}=\tilde{r})\mathsf{P}(R-\tilde{R}=r-\tilde{r})}\\ &=\mathsf{P}((\delta_\pi(G_1,G_2)\leq 0|\tilde{R}=\tilde{r}),\end{aligned} \tag{133}$$

where (133) 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 (128), we have $\tilde{r} = O(\frac{\tilde{t}^{\mathrm{u}}\log n}{n})$ and this inequality follows from (103) from Step 1. Equation (129) follows from the definition of the probability generating function for $\mathrm{Hyp}(r,t^{\mathrm{u}},\tilde{t}^{\mathrm{u}})$. (130) follows form the conclusion about probability generating function of the hypergeometric distribution in (125) with $z_4 \in (0,1)$. (131) is true since $z_4 = O\left(\frac{1}{\log n}\right)$ and $z_5 = O(1)$. (132) is true since $r = O(n\log n)$.

- For the second term of (126), we have

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}>\tilde{r}^* \mid R=r)\\ &=\sum_{\tilde{r}>\tilde{r}^*}\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}=\tilde{r}\mid R=r)\\ &=\sum_{\tilde{r}>\tilde{r}^*}\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0\mid \tilde{R}=\tilde{r})\mathsf{P}(\tilde{R}=\tilde{r}\mid R=r)\end{aligned} \tag{134}$$

$$\leq \max_{0\leq\tilde{r}\leq n}\{\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0|\tilde{R}=\tilde{r})\}\mathsf{P}(\tilde{R}>\tilde{r}^*|R=r).$$

Here (134) 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 (107), 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

$$\begin{aligned}&\max_{0\leq\tilde{r}\leq n}\{\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0\mid\tilde{R}=\tilde{r})\}\\ &\leq \mathsf{P}(\delta_\pi(G_1,G_2)\leq 0\mid\tilde{R}=0)\\ &\leq z_3^{\tilde{n}} z_5^{\tilde{n}},\end{aligned} \tag{135}$$

where (135) follows from (103) in Step 1 with $\tilde{r}=0$. Now we get

$$\begin{aligned}&\mathsf{P}(\delta_\pi(G_1,G_2)\leq 0,\tilde{R}>\tilde{r}^* \mid R=r)\\ &\leq z_3^{\tilde{n}} z_5^{\tilde{n}}\mathsf{P}(\tilde{R}>\tilde{r}^*\mid R=r)\\ &= z_3^{\tilde{n}} z_5^{\tilde{n}}\sum_{i>\tilde{r}^*}[z^i]\Phi_{\mathrm{Hyp}}(z)\end{aligned} \tag{136}$$

$$\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*}\Phi_{\mathrm{Hyp}}(z) \tag{137}$$

$$\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*}\Phi_{\mathrm{Bin}}(z) \tag{138}$$

$$= z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*}\left(1+\tfrac{\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(z-1)\right)^r \tag{139}$$

$$\leq z_3^{\tilde{n}} z_5^{\tilde{n}} z^{-\tilde{r}^*}\exp\left\{\tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(z-1)\right\} \tag{140}$$

$$= z_3^{\tilde{n}} z_5^{\tilde{n}}\exp\left\{-\tilde{r}^*+\tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(e-1)\right\} \tag{141}$$

$$= z_3^{\tilde{n}} z_5^{\tilde{n}}\exp\left\{\tfrac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}(-C-1+e)\right\} \tag{142}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

$$\leq z_3^{\tilde{n}} z_5^{\tilde{n}} \exp\left\{\tfrac{r\tilde{n}(n-2)}{n(n-1)}(-C-1+e)\right\} \tag{143}$$

$$\leq z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(\tfrac{r}{n}(-C-1+e)+O(1)\right)\right\} \tag{144}$$

$$= o\left(z_3^{\tilde{n}} \exp\left\{\tilde{n}\left(-\tfrac{2r}{n}+O(1)\right)\right\}\right) \tag{145}$$

In (136), $\Phi_{\mathrm{Hyp}}(z)$ is a probability generating function for $\mathrm{Hyp}(r, t^{\mathrm{u}}, \tilde{t}^{\mathrm{u}})$. In (137), we set $z > 1$ and the inequality follows from (87) in Fact 3. In (138), $\Phi_{\mathrm{Bin}}(z)$ 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. (139) follows from the definition of $\Phi_{\mathrm{Bin}}(z)$. (140) follows form the inequality $1 + x \leq e^x$. In (141), we set $z = e$. In (142), we plug in $\tilde{r}^* = C\frac{r\tilde{t}^{\mathrm{u}}}{t^{\mathrm{u}}}$ where $C$ is larger than $(e-1)$. In (143), we use the relation $\tilde{t}^{\mathrm{u}} \geq \frac{\tilde{n}(n-2)}{2}$ from (94) and $t^{\mathrm{u}} = \binom{n}{2}$. In (144), we plug in $z_5 = O(1)$. (145) is true because we can always find $C > e+1$ such that (144) is exponentially smaller than (132).

We conclude that the second term of (126) is negligible compared with the upper bound of the first term given in (132). Combining the two terms, (126) can be bounded as

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

*Step 3:* We now establish the desired error bound

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

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

When $nz_6 > 2/3$, we have

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

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

$$\begin{aligned}&\mathsf{P}(\exists \pi \in \mathcal{S}_n \setminus \{\pi_{\mathrm{id}}\}, \delta_\pi(G_1, G_2) \leq 0 \mid R = r)\\ &\leq \sum_{\tilde{n}=2}^{n} \sum_{\pi \in \mathcal{S}_{n,\tilde{n}}} \mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid R = r) &(146)\\ &\leq \sum_{\tilde{n}=2}^{n} |\mathcal{S}_{n,\tilde{n}}| \max_{\pi \in \mathcal{S}_{n,\tilde{n}}} \{\mathsf{P}(\delta_\pi(G_1, G_2) \leq 0 \mid R = r)\}\\ &\leq \sum_{\tilde{n}=2}^{n} |\mathcal{S}_{n,\tilde{n}}| z_6^{\tilde{n}} &(147)\\ &\leq \sum_{\tilde{n}=2}^{n} n^{\tilde{n}} z_6^{\tilde{n}} &(148)\\ &\leq \frac{(nz_6)^2}{1-nz_6}\\ &\leq 3n^2 z_6^2, &(149)\end{aligned}$$

where (146) follows from the union bound, (147) follows from inequality (124) proved in Step 2, (148) follows since $|\mathcal{S}_{n,\tilde{n}}| \leq n^{\tilde{n}}$, and (149) holds since $nz_6 \leq 2/3$.

In summary, $3n^2 z_6^2$ is always an upper bound on the conditional probability. This completes the proof of Lemma 6. □

## VIII. Conclusion and Future Work

In this paper, we proposed the attributed Erdős–Rényi pair model to study the effect of publicly available side information for graph alignment. We established information-theoretic limits for exact alignment, including achievability and converse conditions that match for a certain range of parameters. These conditions can be used to quantify the effect of side information. We also specialized our results to three well-studied graph alignment models for comparison.

There are many more interesting questions to ask about the attributed graph alignment problem. Here we give some example directions. As discussed in Section III, our achievability conditions and converse conditions do not match in the most general scenario. We conjecture that the converse conditions can be potentially improved, especially given the recent developments on tighter converse conditions for the Erdős–Rényi pair model in [8]. Moreover, the achievability results in this work is based on the MAP estimator, which has no polynomial-time implementation. Therefore, a natural question is whether there is any efficient algorithms for attributed graph alignment, and whether there is any fundamental gap between the achievable region by efficient algorithms and the information-theoretic achievable region. This question is partially answered in the concurrent work [12] by proposing two efficient algorithms for attributed graph alignment and analyzing their feasible region. Another direction that is worth further investigation is graph alignment under more general attributed graph models. Our model has assumed that the user-attribute edges are independent of the user-user edges. However, in the social network example, users attending the same institute are more likely to be friends than users attending different institutes. Therefore, it would be interesting to consider graph models in which user-attribute edges are correlated with user-user edges, and to investigate how the correlation affects graph alignment. We comment that a starting point can be the multiplicative attribute graph model proposed in [28], where the probability of a user-user edge depends on the product of individual attribute-attribute similarity.

## Appendix A
## MAP Estimator

In this section, we derive the expression of the MAP for the attributed Erdős–Rényi graph pair model $\mathcal{G}(n, \boldsymbol{p}; m, \boldsymbol{q})$.

*Lemma 7:* 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 $\Pi^*$ 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

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

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

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

*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 $\Pi^*$ can be written as:

$$\begin{aligned}
&\mathsf{P}(\Pi^* = \pi | G_1 = g_1, G_2' = g_2') \\
=&\frac{\mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi)\mathsf{P}(\Pi^* = \pi)}{P(G_1 = g_1, G_2' = g_2')} \\
\propto&\mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi) \qquad (150)\\
=&\mathsf{P}(G_1 = g_1, G_2 = \pi^{-1}(g_2')) \qquad (151)\\
=&\prod_{(i,j)\in\{0,1\}^2} p_{ij}^{\mu_{ij}(g_1, \pi^{-1}(g_2'))} q_{ij}^{\nu_{ij}(g_1, \pi^{-1}(g_2'))}. \qquad (152)
\end{aligned}$$

Here equation (150) follows from the fact that $\Pi^*$ is uniformly drawn from $\mathcal{S}_n$ and $\mathsf{P}(G_1 = g_1, G_2' = g_2')$ does not depend on $\pi$. Equation (151) is due to the independence between $\Pi^*$ and $(G_1, G_2)$.

To further simplify equation (152), 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 the following.

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

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 $\pi$, we can further simplify the posterior as follows

$$\begin{aligned}
&\mathsf{P}(\Pi^* = \pi | G_1 = G_1, G_2' = G_2') \\
\propto&\prod_{(i,j)\in\{0,1\}^2} p_{ij}^{\mu_{ij}(G_1, \pi^{-1}(G_2'))} q_{ij}^{\nu_{ij}(G_1, \pi^{-1}(G_2'))} \\
\propto&\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)^{-\frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2}} \left(\frac{q_{11}q_{00}}{q_{10}q_{01}}\right)^{-\frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2}} \qquad (153)\\
=&\exp\left\{-w_1\frac{\Delta^{\mathrm{u}}(G_1, \pi^{-1}(G_2'))}{2} - w_2\frac{\Delta^{\mathrm{a}}(G_1, \pi^{-1}(G_2'))}{2}\right\}, \qquad (154)
\end{aligned}$$

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 of $G_1$ and $G_2$ are positively correlated. Therefore, of all the permutations 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. □

## Appendix B
## Proof of Lemma 1

Recall that when we specialize the attributed Erdős–Rényi pair model by setting $\boldsymbol{p} = \boldsymbol{q}$, we can treat 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 for the seeded Erdős–Rényi pair—the MAP estimator—also simplifies to minimizing the Hamming distance of the user-user edges and user-seed edges.

*Lemma 8:* Let $(G_1, G_2')$ be a pair of graphs generated from the seeded Erdős–Rényi pair $\mathcal{G}(n, m, \boldsymbol{p})$. The MAP estimator of the permutation $\Pi^*$ based on $(G_1, G_2')$ simplifies to

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

where

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

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

$$\begin{aligned}
&\mathsf{P}(\Pi^* = \pi | G_1 = g_1, G_2' = g_2') \\
&= \frac{\mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi)\mathsf{P}(\Pi^* = \pi)}{P(G_1 = g_1, G_2' = g_2')} \\
&\propto \mathsf{P}(G_1 = g_1, G_2' = g_2' | \Pi^* = \pi) \qquad (155)\\
&= \mathsf{P}(G_1 = g_1, G_2 = \pi^{-1}(g_2')). \qquad (156)
\end{aligned}$$

Here (155) follows since $\Pi^*$ is uniformly drawn. (156) follows since $\Pi^*$ is independent of $G_1$ and $G_2$. For ease of notation, we use $g_2^{\pi}$ to denote $\pi^{-1}(g_2')$. Then according to the seeded graph model in Section II, we have

$$\begin{aligned}
&\mathsf{P}(G_1 = g_1, G_2 = g_2^{\pi}) \\
=&p_{11}^{\mu_{11}(g_1, g_2^{\pi})} p_{00}^{\mu_{00}(g_1, g_2^{\pi})} p_{10}^{\mu_{10}(g_1, g_2^{\pi})} p_{01}^{\mu_{01}(g_1, g_2^{\pi})}. \qquad (157)
\end{aligned}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

In (157), we define

$$\mu_{11}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}$$

$$\mu_{10}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\}$$

$$\mu_{01}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\}$$

$$\mu_{00}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\}.$$

where $\mathcal{V}^{u\prime} \triangleq \mathcal{V}_{\mathrm{u}} \setminus \mathcal{V}^s$ is the set of unmatched user 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

$$\mu'_{11}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\sim} j\right\}$$

$$\mu'_{10}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\}$$

$$\mu'_{01}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\sim} j\right\}$$

$$\mu'_{00}(g_1,g_2^\pi) \triangleq \sum_{i,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i,j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^{u\prime},j\in\mathcal{V}^s} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\} + \sum_{i\in\mathcal{V}^s,j\in\mathcal{V}^{u\prime}} \mathbb{1}\left\{i \overset{g_1}{\not\sim} j, i \overset{g_2^\pi}{\not\sim} j\right\}.$$

We therefore have

$$\mathsf{P}(G_1 = g_1, G_2 = g_2^\pi) \propto p_{11}^{\mu'_{11}(g_1,g_2^\pi)} p_{00}^{\mu'_{00}(g_1,g_2^\pi)} p_{10}^{\mu'_{10}(g_1,g_2^\pi)} p_{01}^{\mu'_{01}(g_1,g_2^\pi)} \tag{158}$$

So far the MAP estimator we derived here is exactly the same as the estimator for attributed graph alignment. Applying Lemma 4, we then get

$$\hat{\pi}_{\mathrm{MAP}}(g_1,g_2') = \operatorname*{argmin}_{\pi\in\mathcal{S}_{\mathrm{u}}} \{\mu'_{10}(g_1,g_2^\pi) + \mu'_{01}(g_1,g_2^\pi)\}.$$

□

## APPENDIX C
## PROOF OF LEMMA 2

*Lemma 9:* 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 $\frac{1}{|\mathrm{Aut}(G_1\wedge G_2)|}$.

*Proof:* We assume without loss of generality that the true underlying permutation $\Pi^*$ us the identity permutation, i.e., $G_2 = G_2'$. Recall that from Lemma 4 we have

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

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

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

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

It suffices to show that for any $\sigma \in \mathrm{Aut}(G_1 \wedge G_2)$, we have $\Delta^{\mathrm{u}}(G_1,G_2) = \Delta^{\mathrm{u}}(G_1,\sigma(G_2))$ and $\Delta^{\mathrm{a}}(G_1,G_2) = \Delta^{\mathrm{a}}(G_1,\sigma(G_2))$. This would imply that permutation $\sigma^{-1}$ has the same posterior as the identity permutation, and thus the estimator cannot do better than a random guess between these permutations.

We firstly show that $\Delta^{\mathrm{u}}(G_1,G_2) = \Delta^{\mathrm{u}}(G_1,\sigma(G_2))$. Consider a user pair $(i,j) \in \mathcal{E}_{\mathrm{u}}$. Suppose that $G_1(i,j) = G_2(i,j) = 1$, i.e., $(G_1 \wedge G_2)(i,j) = 1$. Because $\sigma \in \mathrm{Aut}(G_1 \wedge G_2)$, we know that $G_1(\sigma(i),\sigma(j)) = G_2(\sigma(i),\sigma(j)) = 1$. So the contribution of $(i,j)$ to $\Delta^{\mathrm{u}}(G_1,G_2)$ and $\Delta^{\mathrm{u}}(G_1,\sigma(G_2))$ are both zero. Next we consider the case of $(G_1 \wedge G_2)(i,j) = 0$. This includes subcases of $G_1(i,j) = 0, G_2(i,j) = 1$ and $G_1(i,j) = 1, G_2(i,j) = 0$ and $G_1(i,j) = 0, G_2(i,j) = 0$. Let $S_{\mathrm{u}}$ denote the edge orbit in the permutation $\sigma^{\mathcal{E}}$ that contains $(i,j)$. Because $\sigma \in \mathrm{Aut}(G_1 \wedge G_2)$, we know that $(G_1 \wedge G_2)(e) = 0$ for each $e \in S_{\mathrm{u}}$. Note that the contribution of $e$ to $\Delta^{\mathrm{u}}(G_1,G_2)$ is one if $G_1(e) = 0, G_2(e) = 1$ or $G_1(e) = 1, G_2(e) = 0$ and the contribution is zero if $G_1(e) = 0, G_2(e) = 0$. Therefore, the total contribution of the orbit $S_{\mathrm{u}}$ to $\Delta^{\mathrm{u}}(G_1,G_2)$ is given by the total number of edges in $G_1$ and $G_2$ on the orbit $S_{\mathrm{u}}$. Similarly, the contribution of $S_{\mathrm{u}}$ to $\Delta^{\mathrm{u}}(G_1,\sigma(G_2))$ is the same. Thus, we have $\Delta^{\mathrm{u}}(G_1,G_2) = \Delta^{\mathrm{u}}(G_1,\sigma(G_2))$.

Secondly, we show that $\Delta^{\mathrm{a}}(G_1,G_2) = \Delta^{\mathrm{a}}(G_1,\sigma(G_2))$. Consider a user-attribute pair $(i,a) \in \mathcal{E}_{\mathrm{a}}$. Then if $(G_1 \wedge G_2)(i,a) = 1$, we have $G_1(i,a) = G_2(i,a) = G_1(\sigma(i),a) = G_2(\sigma(i),a) = 1$. So the contribution of $(i,a)$ to $\Delta^{\mathrm{a}}(G_1,G_2)$ and $\Delta^{\mathrm{a}}(G_1,\sigma(G_2))$ are both zero. Next, suppose $(G_1 \wedge G_2)(i,a) = 0$. Let $S_{\mathrm{a}}$ denote the edge orbit that contains $(i,a)$. It is not hard to see that the contribution of the user-attribute pairs in $S_{\mathrm{a}}$ to $\Delta^{\mathrm{a}}(G_1,G_2)$ equals to the total number of edges

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

in $G_1$ and $G_2$ on $S_{\mathrm{a}}$, and its contribution of $\Delta^{\mathrm{a}}(G_1, \sigma(G_2))$ is the same. So we have $\Delta^{\mathrm{a}}(G_1, G_2) = \Delta^{\mathrm{a}}(G_1, \sigma(G_2))$. □

## APPENDIX D
## PROOF OF FACT 1

*Fact 4:* The generating function $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z)$ of permutation $\pi$ 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 $t_l^{\mathrm{u}}$ is the number of user-user orbits of size $l$, $t_l^{\mathrm{a}}$ 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 $\pi$

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

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 $i$th orbit and there are $N$ orbits in total. Then we have the following.

$$\begin{aligned}
&\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}, z) \\
&= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} z^{\delta_\pi(g,h)} \boldsymbol{x}^{\boldsymbol{\mu}(g,h)} \boldsymbol{y}^{\boldsymbol{\nu}(g,h)} \\
&= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} \prod_{e \in \mathcal{E}} z^{\delta_\pi(g_e,h_e)} \boldsymbol{x}^{\boldsymbol{\mu}(g_e,h_e)} \boldsymbol{y}^{\boldsymbol{\nu}(g_e,h_e)} && (159) \\
&= \sum_{g \in \{0,1\}^{\mathcal{E}}} \sum_{h \in \{0,1\}^{\mathcal{E}}} \prod_{i=1}^{N} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) && (160) \\
&= \sum_{g_{\mathcal{O}_1} \in \{0,1\}^{\mathcal{O}_1}} \sum_{h_{\mathcal{O}_1} \in \{0,1\}^{\mathcal{O}_1}} \cdots \sum_{h_{\mathcal{O}_N} \in \{0,1\}^{\mathcal{O}_N}} \prod_{i=1}^{N} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) && (161) \\
&= \prod_{i=1}^{N} \left( \sum_{g_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} \sum_{h_{\mathcal{O}_i} \in \{0,1\}^{\mathcal{O}_i}} f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i}) \right) && (162) \\
&= \prod_{i=1}^{N} \mathcal{A}_{\mathcal{O}_i}(\boldsymbol{x}, \boldsymbol{y}, z) && (163) \\
&= \prod_{l \geq 1} \mathcal{A}_l(\boldsymbol{x}, z^{w_1})^{t_l^{\mathrm{u}}} \mathcal{A}_l(\boldsymbol{y}, z^{w_2})^{t_l^{\mathrm{a}}}. && (164)
\end{aligned}$$

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 (159), $g_e$ (resp. $h_e$) represent a subset of $g$ (resp. $h$) that contains a single vertex pair $e$. In (160), $g_{\mathcal{O}_i}$ (resp. $h_{\mathcal{O}_i}$) represents the subset of $g$ (resp. $h$) that contains only vertex pairs in the 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 (161) 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, (162) follows because $f(g_{\mathcal{O}_i}, h_{\mathcal{O}_i})$'s are independent functions. In (163), 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 why 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 (164), we apply the fact that orbits of the same size have the same generating function. □

## APPENDIX E
## A USEFUL FACT FOR COROLLARIES 1 AND 2

*Fact 5:* Consider the subsampling representation of the graph parameters

$$\begin{aligned}
&\begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} \\
&= \begin{pmatrix} p s_{\mathrm{u},1} s_{\mathrm{u},2} & p s_{\mathrm{u},1}(1 - s_{\mathrm{u},2}) \\ p(1 - s_{\mathrm{u},1}) s_{\mathrm{u},2} & p(1 - s_{\mathrm{u},1})(1 - s_{\mathrm{u},2}) + 1 - p \end{pmatrix},
\end{aligned}$$

If $1 - p = \Theta(1)$, then we have $\psi_{\mathrm{u}} = \Theta(p_{11})$ and $\psi_{\mathrm{u}} = p_{11} - \Theta(p_{11} p^{1/2})$. The statement holds if we exchange $\boldsymbol{p}$ to $\boldsymbol{q}$.

*Proof:* To see $\psi_{\mathrm{u}} = \Theta(p_{11})$, we write $\psi_{\mathrm{u}}$ using parameters from the subsampling model and we have

$$\begin{aligned}
\psi_{\mathrm{u}} &= (\sqrt{p_{11} p_{00}} - \sqrt{p_{10} p_{01}})^2 \\
&= (\sqrt{p_{11}((1-p) + p(1 - s_{\mathrm{u},1})(1 - s_{\mathrm{u},2}))} \\
&\quad - \sqrt{p^2 s_{\mathrm{u},1} s_{\mathrm{u},2}(1 - s_{\mathrm{u},1})(1 - s_{\mathrm{u},2})})^2 \\
&= (1-p) p_{11} \\
&\quad \cdot \left( \sqrt{1 + \tfrac{p(1-s_{\mathrm{u},1})(1-s_{\mathrm{u},2})}{1-p}} - \sqrt{\tfrac{p(1-s_{\mathrm{u},1})(1-s_{\mathrm{u},2})}{1-p}} \right)^2 \\
&= \frac{(1-p) p_{11}}{\left( \sqrt{1 + \tfrac{p(1-s_{\mathrm{u},1})(1-s_{\mathrm{u},2})}{1-p}} + \sqrt{\tfrac{p(1-s_{\mathrm{u},1})(1-s_{\mathrm{u},2})}{1-p}} \right)^2}.
\end{aligned} \tag{165}$$

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

In (165), we have that $(1-p) = \Theta(1)$ and $\frac{1}{\sqrt{1+\frac{p(1-s_{u,1})(1-s_{u,2})}{1-p}}+\sqrt{\frac{p(1-s_{u,1})(1-s_{u,2})}{1-p}}} = \Theta(1)$. Therefore, $\psi_u = \Theta(p_{11})$.

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

$$\begin{aligned}\psi_u &= (\sqrt{p_{11}p_{00}} - \sqrt{p_{10}p_{01}})^2 \\ &= p_{11}p_{00} + p_{10}p_{01} - 2\sqrt{p_{11}p_{00}p_{10}p_{01}} \\ &= p_{11}((1-p) + p(1-s_{u,1})(1-s_{u,2})) \\ &\quad + p^2 s_{u,1}s_{u,2}(1-s_{u,1})(1-s_{u,2}) \\ &\quad - \sqrt{p_{11}^2((1-p) + p(1-s_{u,1})(1-s_{u,2}))p(1-s_{u,1})(1-s_{u,2})} \\ &= p_{11} - O(p_{11}p^{1/2}).\end{aligned}$$

□

## APPENDIX F
## PROOF OF COROLLARY 1

### A. Achievability

In this proof, we first show that, under the assumptions on the user-user edges in condition (12) and (14), the achievability result becomes

$$np_{11} + m\psi_a - \log n = \omega(1).$$

Next, we apply conditions (13) and (15) to bound difference between $q_{11}$ and $\psi_a$, and complete the proof.

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 (12), we also have $\psi_u = \omega(\frac{\log n}{n})$ because $\psi_u = \Theta(p_{11})$ (see Fact 5 in Appendix E). Therefore exact alignment is achievable according to Theorem 1: $\frac{n\psi_u}{2} + m\psi_a - \log n = \omega(\log n) + m\psi_a - \log n = \omega(1)$. Now we check the case when $p_{11} = O(\frac{\log n}{n})$. We will see that all the conditions in Theorem 2 are satisfied. Notice that $p_{10} = ps_{u,1}(1-s_{u,2}) \le ps_{u,1} = \frac{p_{11}}{s_{u,2}}$ and $p_{01} = ps_{u,2}(1-s_{u,1}) \le ps_{u,2} = \frac{p_{11}}{s_{u,1}}$. Under assumption (14), we know that $s_{u,1} = \Omega(\frac{(\log n)^4}{n})$ and $s_{u,2} = \Omega(\frac{(\log n)^4}{n})$. Because $p_{11} = O(\frac{\log n}{n})$, we have $p_{01} = O(\frac{1}{\log n})$ and $p_{10} = O(\frac{1}{\log n})$. Moreover, note that

$$\frac{p_{10}p_{01}}{p_{00}p_{11}} = O\left(\frac{p^2 s_{u,1}s_{u,2}(1-s_{u,1})(1-s_{u,2})}{ps_{u,1}s_{u,2}}\right) = O(p).$$

Because $ps_{u,1}s_{u,2} = O\left(\frac{\log n}{n}\right)$ and $s_{u,1}s_{u,2} = \Omega\left(\frac{(\log n)^4}{n}\right)$, we have $p = O\left(\frac{1}{(\log n)^3}\right)$. Then it follows that the sparsity conditions in Theorem 2 (7), (8) and (9) are all satisfied. Therefore, we just need $np_{11} + m\psi_a - \log n = \omega(1)$ to guarantee that exact alignment is achievable. Combining the two case, we come to the conclusion that, under the assumptions (12) (14), the achievability results in Theorem 1 and Theorem 2 simplifies to

$$np_{11} + m\psi_a - \log n = \omega(1). \tag{166}$$

From the above discussion, we further simplify the achievability results so that it depends only on $np_{11}$ and $mq_{11}$, which are the two only parameters in the converse bound. Then we can show in what regime the achievability and converse are tight (up to $\pm\omega(1)$). From the achievability in last step: $np_{11} + m\psi_a - \log n = \omega(1)$, we then need to determine the difference between $m\psi_a$ and $mq_{11}$. Firstly, consider the case when $mq_{11} = \omega(\log n)$. In this case, we immediately have that $np_{11} + m\psi_a - \log n = \omega(1)$ because $\psi_a = \Theta(q_{11})$ by Fact 5. Now, consider the case when $mq_{11} = O(\log n)$. Suppose $np_{11} + mq_{11} - \log n = \omega(1)$ implies $mq_{11}q^{1/2} = O(1)$, then it implies $np_{11} + m\psi_a - \log n = \omega(1)$ as well. This is because $mq_{11} - m\psi_a = O(mq_{11}q^{1/2})$. Therefore, we need to find the condition for $np_{11} + mq_{11} - \log n = \omega(1)$ to imply $mq_{11}q^{1/2} = O(1)$. Because $mq_{11} = mqs_{a,1}s_{a,2} = O(\log n)$, we have $mq_{11}q^{1/2} = mq_{11}\left(\frac{mq_{11}}{ms_{a,1}s_{a,2}}\right)^{1/2} = O\left(\frac{(\log n)^{3/2}}{(ms_{a,1}s_{a,2})^{1/2}}\right)$. Condition (15) implies that $ms_{a,1}s_{a,2} = \Omega((\log n)^3)$, so we have that $mq_{11}q^{1/2} = O(1)$, which completes the proof.

### B. Converse

In this proof, we will show that

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

implies condition (11) in Theorem 3. Note that $-\log(x^2 + (1-x)^2) \ge 2x$ for any $x \in [0,1]$. Therefore, we have

$$\begin{aligned}&-n\log(p_{11}^2 + (1-p_{11})^2) - m\log(q_{11}^2 + (1-q_{11})^2) \\ &\le 2np_{11} + 2mq_{11} \le 2\log n - \omega(1),\end{aligned}$$

which completes the proof.

## APPENDIX G
## PROOF OF COROLLARY 2

### A. Proof for the Achievability Condition (41)

Recall that from (166) in the proof of Corollary 1, we have, under assumptions (12) and (14), our achievability results (Theorem 1 and Theorem 2) simplify to the following condition

$$np_{11} + m\psi_a \ge \log n + \infty,$$

where $\psi_a - q_{11} = O(q_{11}q^{1/2})$.

Now, in the seeded Erdős–Rényi setting, we have $\boldsymbol{p} = \boldsymbol{q}$. Because condition (31) implies conditions (12) and (14), we obtain the following achievability for seeded alignment

$$np_{11} + m\psi_u \ge \log n + \omega(1), \tag{167}$$

where $\psi_u - p_{11} = O(p_{11}p^{1/2})$.

For the above achievability condition (167), we show that it is equivalent to

$$(n+m)p_{11} \ge \log n + \omega(1) \tag{168}$$

by comparing them in the following three regimes.

1) For the regime $(n+m)p_{11} = \omega(\log n)$, we show that it is strictly contained in both (167) and (168). We can easily see that $(n+m)p_{11} = \omega(\log n)$ satisfy condition (168). For condition (167), recall that we have $\psi_u = \Theta(p_{11})$ from Fact 5 (cf. Appendix E). Thus, we also have $(n+m)p_{11} = \omega(\log n)$ satisfy condition (168).

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

2) In the regime $(n+m)p_{11} = \Theta(\log n)$, we have

$$\begin{aligned} mp_{11}p^{1/2} &= O\left(mp_{11}\left(\frac{\log n}{(m+n)s^2}\right)^{1/2}\right) \\ &= O\left(\log n\left(\frac{\log n}{ns^2}\right)^{1/2}\right) \\ &= O\left(\frac{(\log n)^{3/2}}{(\log n)^2}\right) \\ &= O(1), \end{aligned}$$

where the penultimate equality follows by assumption (14). For the condition (167), we have $m\psi_{\mathrm{u}} = mp_{11} - O(mp_{11}p^{1/2}) = mp_{11} - O(1)$. Therefore, condition (167) can be simplified to $np_{11} + mp_{11} \geq \log n + \omega(1)$, which is exactly condition (168).

3) For the regime $(n+m)p_{11} = o(\log n)$, it is not contained by neither (167) nor (168).

### B. Proof for the Converse Condition (42)

From Theorem 3, we have the converse condition for attributed Erdős–Rényi alignment

$$np_{11} + mq_{11} \leq \log n - \omega(1).$$

Now, in the seeded Erdős–Rényi setting, we have $\boldsymbol{p} = \boldsymbol{q}$ and we directly obtain the following converse for seeded alignment

$$(n+m)p_{11} \leq \log n - \omega(1). \tag{169}$$

## ACKNOWLEDGMENT

The authors would like to express their gratitude to the associate editor and the anonymous reviewers for their invaluable contribution and constructive feedback, which have greatly improved the quality and clarity of this article.

## REFERENCES

[1] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 35, pp. 12763–12768, Sep. 2008.

[2] M. Cho and K. M. Lee, "Progressive graph matching: Making a move of graphs via probabilistic voting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 398–405.

[3] A. D. Haghighi, A. Y. Ng, and C. D. Manning, "Robust textual inference via graph matching," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process. HLT*, 2005, pp. 387–394.

[4] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 173–187.

[5] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proc. VLDB Endowment*, vol. 7, no. 5, pp. 377–388, Jan. 2014.

[6] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 1235–1243.

[7] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for Erdős-Rényi graph matching," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 44, no. 1, pp. 63–72, Jun. 2016.

[8] Y. Wu, J. Xu, and S. H. Yu, "Settling the sharp reconstruction thresholds of random graph matching," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5391–5417, Aug. 2022.

[9] D. Cullina and N. Kiyavash, "Exact alignment recovery for correlated Erdős-Rényi graphs," 2017, *arXiv:1711.06783*.

[10] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy*, 2008, pp. 111–125.

[11] E. Mossel and J. Xu, "Seeded graph matching via large neighborhood statistics," *Random Struct. Algorithms*, vol. 57, no. 3, pp. 570–611, Oct. 2020.

[12] Z. Wang, N. Zhang, W. Wang, and L. Wang, "On the feasible region of efficient algorithms for attributed graph alignment," *IEEE Trans. Inf. Theory*, vol. 70, no. 5, pp. 3622–3639, May 2024.

[13] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 651–655.

[14] O. E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser, "Analysis of a canonical labeling algorithm for the alignment of correlated Erdős-Rényi graphs," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 3, no. 2, pp. 1–25, Jun. 2019.

[15] J. Ding, Z. Ma, Y. Wu, and J. Xu, "Efficient random graph matching via degree profiles," *Probab. Theory Rel. Fields*, vol. 179, nos. 1–2, pp. 29–115, Feb. 2021.

[16] Z. Fan, C. Mao, Y. Wu, and J. Xu, "Spectral graph matching and regularized quadratic relaxations: Algorithm and theory," in *Proc. 37th Int. Conf. Mach. Learn.*, H. D. III and A. Singh, Eds., vol. 119, Jul. 2020, pp. 2985–2995.

[17] C. Mao, M. Rudelson, and K. Tikhomirov, "Exact matching of random graphs with constant correlation," *Probab. Theory Rel. Fields*, vol. 186, nos. 1–2, pp. 327–389, Jun. 2023.

[18] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Random graph matching at otter's threshold via counting chandeliers," in *Proc. 55th Annu. ACM Symp. Theory Comput.*, 2023, pp. 1345–1356.

[19] V. Lyzinski, D. E. Fishkind, and C. E. Priebe, "Seeded graph matching for correlated Erdős-Rényi graphs," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3513–3540, 2014.

[20] D. E. Fishkind et al., "Seeded graph matching," *Pattern Recognit.*, vol. 87, pp. 203–215, Mar. 2019.

[21] S. Zhang and H. Tong, "Attributed network alignment: Problem definitions and fast solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1680–1692, Sep. 2019.

[22] S. Zhang and H. Tong, "FINAL: Fast attributed network alignment," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1345–1354.

[23] Q. Zhou, L. Li, X. Wu, N. Cao, L. Ying, and H. Tong, "Attent: Active attributed network alignment," in *Proc. Web Conf.*, Apr. 2021, pp. 3896–3906.

[24] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching," in *Proc. 1st ACM Conf. Online social Netw.*, Oct. 2013, pp. 119–130.

[25] J. Lubars and R. Srikant, "Correcting the output of approximate graph matching algorithms," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2018, pp. 1745–1753.

[26] J. Edmonds and R. M. Karp, "Theoretical improvements in algorithmic efficiency for network flow problems," *J. ACM*, vol. 19, no. 2, pp. 248–264, Apr. 1972.

[27] V. Chvátal, "The tail of the hypergeometric distribution," *Discrete Math.*, vol. 25, no. 3, pp. 285–287, 1979.

[28] M. Kim and J. Leskovec, "Multiplicative attribute graph model of real-world networks," *Internet Math.*, vol. 8, nos. 1–2, pp. 113–160, Mar. 2012.

**Ning Zhang** received the bachelor's degree in physics from Nankai University in 2019 and the M.A.Sc. degree in electrical and computer engineering from The University of British Columbia in 2022. She is currently pursuing the D.Phil. degree in computational discovery with the Department of Statistics, University of Oxford. Her research interests include the intersection of statistics and computation, with a current focus on graph-based machine learning and signal processing.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.

**Ziao Wang** (Graduate Student Member, IEEE) received the B.E. degree in electrical and computer engineering from Nanyang Technological University in 2019 and the M.A.Sc. degree in electrical and computer engineering from The University of British Columbia, Vancouver, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include information theory, statistical inference on graphs, and noisy computing.

**Weina Wang** (Member, IEEE) received the bachelor's degree from the Department of Electronic Engineering, Tsinghua University, in 2009, and the Ph.D. degree in electrical engineering from Arizona State University in 2016. She was a Joint Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, and the School of ECEE, Arizona State University, from 2016 to 2018. She is currently an Assistant Professor with the Computer Science Department, Carnegie Mellon University. Her research interests include applied probability and stochastic systems, with applications in cloud computing, data centers, and privacy-preserving data analytics. Her dissertation received the Dean's Dissertation Award from the Ira A. Fulton Schools of Engineering, Arizona State University, in 2016. She received the Kenneth C. Sevcik Outstanding Student Paper Award from ACM SIGMETRICS 2016, the Best Paper Award from ACM MobiHoc 2022, the NSF CAREER Award in 2022, and the ACM SIGMETRICS Rising Star Research Award in 2023.

**Lele Wang** (Member, IEEE) received the B.E. degree from Tsinghua University, China, in 2009, and the Ph.D. degree in communication theory and systems from the University of California at San Diego, San Diego, CA, USA, in 2015. She attended the Academic Talent Program. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of British Columbia (UBC), Vancouver. Before joining UBC, she was an NSF Center for Science of Information Post-Doctoral Fellow. From 2015 to 2017, she was a Post-Doctoral Researcher with Stanford University and Tel Aviv University. Her research interests include information theory, coding theory, communication theory, statistical inference on graphs, high dimensional statistics, and generative models in machine learning. She was a recipient of the 2013 UCSD Shannon Memorial Fellowship, the 2013–2014 Qualcomm Innovation Fellowship, and the 2017 NSF Center for Science of Information Postdoctoral Fellowship. Her Ph.D. thesis "Channel Coding Techniques for Network Communication" won the 2017 IEEE Information Theory Society Thomas M. Cover Dissertation Award. She is one of the tutorial speakers at IEEE ISIT 2024. She serves as the Vice President for the Canadian Society of Information Theory during 2022–2025, an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS during 2023–2026, and the Guest Associate Editor for IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY in 2023.

Authorized licensed use limited to: Carnegie Mellon University Libraries. Downloaded on April 01,2025 at 20:53:16 UTC from IEEE Xplore. Restrictions apply.