Correlated Stochastic Block Models: Exact Graph Matching with Applications to Recovering Communities

Anonymous Author(s)
Affiliation
Address
email

Abstract

Motivated by applications in machine learning, we consider the task of learning latent community structure from multiple correlated networks. First, we study the problem of learning the latent vertex correspondence between two edge-correlated stochastic block models, focusing on the regime where the average degree is logarithmic in the number of vertices. We derive the precise information-theoretic threshold for exact recovery: above the threshold there exists an estimator that outputs the true correspondence with probability close to 1, while below the threshold no estimator can recover the true correspondence with probability bounded away from 0. As an application of our results, we show how one can exactly recover the latent communities using *multiple* correlated graphs in parameter regimes where it is information-theoretically impossible to do so using just a single graph.

1 Introduction

Learning community structure in networks is a ubiquitous inference task in several domains, including biology [11, 36], sociology [23], and machine learning [52, 33, 54]. The last few decades have therefore seen an explosion of work on the topic, leading to determining the fundamental information-theoretic limits for learning communities in probabilistic generative models [1, 3, 2, 40], as well as algorithms that work well in practice [50, 30, 19]. Typically, such algorithms only leverage the structure of the network (i.e., the configuration of node-node links). Increasingly, one often has access to *side information* that can greatly improve the performance of inference algorithms.

There is a vast literature on designing algorithms that incorporate various types of side information to aid in recovering communities in networks. The works [18, 42, 29, 51, 34] leverage *node-level* information that is correlated with community memberships. Another line of work [5, 47, 48, 32, 4, 7, 38, 34] recovers communities from a *multi-layer* network, where the distribution of edges in each layer is influenced by the same underlying community structure.

In contrast to prior work, we explore scenarios where the side information comes in the form of *multiple correlated networks*, which is natural in several domains including social networks [44, 49, 31], bioinformatics [53], and machine learning [13, 12]. In the context of social networks, for instance, many datasets are anonymized to protect the identity of users. Nevertheless, one may be able to infer additional information about users from additional networks by noting that the interaction patterns of the same set of users are likely to be similar across networks [44, 31]. In bioinformatics, an important goal is to study the functional properties of proteins groups through a protein-protein interaction (PPI) network. Using the insight that functionally similar protein groups will have similar interaction structures, one can compare PPIs across species to infer protein functions [53]. In all

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

of these examples, an important task, commonly known as *graph matching*, is to synthesize the information from multiple correlated networks in a sensible manner.

To the best of our knowledge, we are the first to consider the use of multiple correlated networks for recovering communities. Specifically, we quantify, in an information-theoretic sense, how much information we can gain from correlated networks in order to identify community structure. To this end, we shall focus on correlated networks G_1 and G_2 drawn according to the stochastic block model (SBM), which is widely recognized as the canonical probabilistic generative model for networks with community structure (see, e.g., [1]). The reason for studying this probabilistic model is twofold. For one, it serves as a prototypical model for networks with community structure found in practice, hence the algorithms we develop will serve as a starting point for applications. Moreover, the SBM has well-defined *ground-truth* communities, so we can concretely study the correctness of algorithms in terms of whether the communities they output align with the ground truth.

2 Models and Questions

The SBM is perhaps the simplest and most well-known probabilistic generative model for networks with community structure. It was initially proposed by Holland, Laskey and Leinhardt [28] and subsequently used as a theoretical testbed for evaluating clustering algorithms on average-case networks (see, e.g., [21, 10, 8]). A striking fact about the SBM is that it exhibits sharp information-theoretic phase transitions for various inference tasks, leading to a *precise* understanding of when community information can be extracted from network data. Such phase transitions were first conjectured by Decelle, Krzakala, Moore and Zdeborová [17] and were subsequently made formal by several authors [3, 2, 40, 37, 41, 9, 39, 1]. In summary, the SBM is a well-motivated and mathematically rich setting for studying inference tasks.

In this work we focus on the simplest setting, a SBM with two symmetric communities. For a positive integer n and $p, q \in [0,1]$, we construct $G \sim \mathrm{SBM}(n,p,q)$ as follows. We start with a collection of n vertices, indexed by elements of $[n] := \{1,\ldots,n\}$. We let $\boldsymbol{\sigma} := \{\sigma_i\}_{i=1}^n$ be a vector of community labels where σ_i is a uniform random element of $\{-1,+1\}$, chosen independently across all $i \in [n]$. Based on the community labeling $\boldsymbol{\sigma}$, the two communities are given by the sets $V_+ := \{i \in [n] : \sigma_i = +1\}$ and $V_- := \{i \in [n] : \sigma_i = -1\}$. Then, given the community labels $\boldsymbol{\sigma}$, the edges are drawn independently across vertex pairs as follows. For distinct $i, j \in [n]$, if $\sigma_i \sigma_j = 1$, then the edge (i,j) is in G with probability p; else (i,j) is in G with probability q.

Generally speaking, a community recovery algorithm takes as input G and outputs a community labeling $\widehat{\boldsymbol{\sigma}}$. The *overlap* between the estimated labeling and the ground truth is given by

$$O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) := \frac{1}{n}\left|\sum_{i=1}^n \widehat{\boldsymbol{\sigma}}_i \boldsymbol{\sigma}_i\right|.$$

In the formula for the overlap, we take an absolute value since the labelings $\boldsymbol{\sigma}$ and $-\boldsymbol{\sigma}$ specify the same community partition (and it is only possible to recover $\boldsymbol{\sigma}$ up to its sign). Moreover, notice that $O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma})$ is always between 0 and 1, where a larger value corresponds to a better match between the estimated communities and the ground truth. Indeed, if the algorithm succeeds in exactly recovering the communities (i.e., $\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$ or $\widehat{\boldsymbol{\sigma}} = -\boldsymbol{\sigma}$) then $O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1$.

In the logarithmic degree regime – that is, when $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ for some $\alpha, \beta \geq 0$ – it is well-known that there is a sharp information-theoretic threshold for exactly recovering communities in the SBM [3, 2, 40, 1]. If

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| \geq \sqrt{2}, \tag{1}$$

then exact recovery is possible: there is a polynomial-time algorithm which outputs an estimator $\widehat{\boldsymbol{\sigma}}$ satisfying $\lim_{n\to\infty} \mathbb{P}(O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$. On the other hand, if

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| < \sqrt{2}, \tag{2}$$

then exact recovery is impossible: for *any* estimator $\widetilde{\boldsymbol{\sigma}}$, $\lim_{n\to\infty} \mathbb{P}(O(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 0$.

The goal of our work is to understand how side information in the form of multiple *correlated* SBMs affects the thresholds (1) and (2). To construct a pair of correlated SBMs, we define an additional

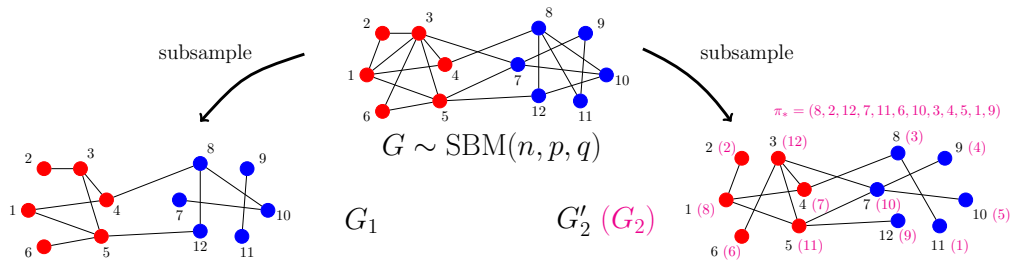

Figure 1: Schematic showing the construction of correlated SBMs (see text for details).

parameter $s \in [0, 1]$ which controls the level of correlation between the two graphs. Formally, we construct $(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$ as follows. First, π_* is chosen to be a uniform random permutation of $[n]$, and then we independently generate a parent graph $G \sim \mathrm{SBM}(n, p, q)$ (note in particular that π_* and the community labelings $\boldsymbol{\sigma}$ are independent). We construct G_1 and G_2' independently conditioned on G by including each edge in G in G_1 with probability s, with the same holding for the construction of G_2'. Finally, we generate G_2 by relabeling the vertices of G_2' according to π_* (e.g., vertex i in G_2' is relabeled to $\pi_*(i)$). This last step in the construction of G_2 reflects the observation that in applications, node labels are often obscured. This construction is visualized in Figure 1.

This model of correlated SBMs was first studied by Onaran, Erkip and Garg [45]. We remark that this process of generating correlated graphs (i.e., by first generating a parent graph, independently subsampling it and randomly permuting the labels) is a common approach for inducing correlation in the formation of edges, and has been employed to study correlated graphs from the Erdős-Rényi model (see, e.g., [49] as well as further references in Section 4), the Chung-Lu model [56] and the preferential attachment model [31].

Since the subsampling probability is s in the construction of the correlated graphs, the marginal distributions of G_1 and G_2 are both $\mathrm{SBM}(n, ps, qs)$. In the logarithmic degree regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, (1) implies that the communities can be exactly recovered from G_1 (or G_2) *alone* if

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| \geq \sqrt{\frac{2}{s}}. \tag{3}$$

A central question of our work is how one can utilize the side information in G_2 to go *beyond* the single-graph threshold (3). This is formalized below.

Objective 2.1 (Exact community recovery)**.** *Given* $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$*, determine conditions on* α, β, s *so that there exists an estimator* $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ *satisfying*

$$\lim_{n \to \infty} \mathbb{P}(O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1.$$

A key observation is that if the latent correspondence π_* is known, one can readily improve the achievability region in (3). Indeed, if π_* is known, one can reconstruct G_2' from G_2. We can then construct a new graph H_* by "overlaying" G_1 and G_2' (i.e., taking their union). Formally, (i, j) is an edge in H_* if and only if (i, j) is an edge in G_1 or G_2'. An equivalent interpretation is that (i, j) is an edge in the parent graph G and that the edge is included in either G_1 or G_2' in the subsampling process. The probability that the edge is *not* included in both G_1 or G_2' is $(1 - s)^2$, so it follows that

$$H_* \sim \mathrm{SBM}\left(n, \alpha(1 - (1 - s)^2)\frac{\log n}{n}, \beta(1 - (1 - s)^2)\frac{\log n}{n}\right).$$

Applying a community recovery algorithm to H_* shows that exact community recovery is possible if

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| \geq \sqrt{\frac{2}{1 - (1 - s)^2}}. \tag{4}$$

Since $1 - (1 - s)^2 > s$ for $s \in (0, 1)$, (4) is a strict improvement over (3). Remarkably, this implies that if π_* is known and if

$$\sqrt{\frac{2}{s}} > \left|\sqrt{\alpha} - \sqrt{\beta}\right| \geq \sqrt{\frac{2}{1 - (1 - s)^2}},$$

then it is information-theoretically impossible to exactly recover $\boldsymbol{\sigma}$ from G_1 or G_2 alone, but one can recover $\boldsymbol{\sigma}$ exactly by combining information from G_1 *and* G_2. To make this analysis rigorous, we study when it is possible to exactly recover π_* from G_1, G_2. We refer to this task as *graph matching*.

Objective 2.2 (Exact graph matching)**.** *Given* $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$*, determine conditions on* α, β, s *so that there exists an estimator* $\widehat{\pi} = \widehat{\pi}(G_1, G_2)$ *satisfying*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\pi} = \pi_*) = 1.$$

While we have motivated graph matching as an intermediate step in recovering communities, it is an important problem in its own right, with applications to data privacy in social networks [44, 49], protein-protein interaction networks [53], and machine learning [13, 12], among others. In particular, it is well known that graph matching algorithms can be used to de-anonymize social networks [44], showing that anonymity is not the same as privacy. Studying the fundamental limits of when graph matching is possible can serve to highlight the precise conditions when anonymity can indeed guarantee privacy and when additional safeguards are necessary.

Although Objective 2.2 has not been previously investigated in the literature, there is strong evidence that there is a phase transition for exact recovery of π_* in the logarithmic degree regime. In the special case of correlated Erdős-Rényi graphs—that is, when $\alpha = \beta$—the maximum likelihood estimate exactly recovers π_* with probability tending to 1 if $s^2\alpha > 1$. When $\alpha \neq \beta$ and $\boldsymbol{\sigma}$ is known, Onaran, Garg and Erkip [46] showed that exact recovery of π_* is possible if

$$s\left(1 - \sqrt{1 - s^2}\right)\left(\frac{\alpha + \beta}{6}\right) > 1. \tag{5}$$

Since the analysis in [46] relies on knowledge of $\boldsymbol{\sigma}$, it is unclear if (5) implies that one can estimate π_* based on knowledge of G_1 and G_2 only. Nevertheless, their result provides further evidence that exact graph matching may be possible in the logarithmic degree regime.

Impossibility results for exact graph matching in correlated SBMs have not been previously determined. In correlated Erdős-Rényi graphs, if $s^2\alpha < 1$, it is known that there is no estimator which exactly recovers π_* with probability bounded away from zero [15, 14, 55]. In particular, the information-theoretic threshold $s^2\alpha = 1$ is the *connectivity threshold* for the intersection graphof G_1 and G_2'. (Given two graphs H_1, H_2, the edge (i, j) is in the intersection graph of H_1 and H_2 if and only if it is an edge in both H_1 and H_2.) For G_1 and G_2' generated through the correlated SBM distribution, the connectivity threshold for the intersection graph is

$$s^2\left(\frac{\alpha + \beta}{2}\right) = 1. \tag{6}$$

Generalizing ideas from the literature on the correlated Erdős-Rényi model suggests that (6) is the correct information-theoretic threshold for exact recovery of π_* as well. In Theorems 3.1 and 3.2, we will show that this is indeed the case.

3 Main results

In this section, we provide answers to the questions posed by Objectives 2.1 and 2.2. In Section 3.1, we precisely characterize the fundamental information-theoretic limits for exact graph matching, thereby fully achieving Objective 2.2. In Section 3.2, we provide partial answers to Objective 2.1; in particular, these provide the fundamental information-theoretic limits for exact community recovery in the regime $s^2(\alpha+\beta)/2 > 1$. Finally, in Section 3.3, we extend the ideas of Section 3.2 to establish achievability and impossibility results for exact community recovery with K correlated SBMs. Phase diagrams illustrating the results can be found in the Supplementary Material.

3.1 Exact Graph Matching

In our first set of results, we derive the precise information-theoretic threshold, above which exact graph matching is possible and below which no estimator can output the correct vertex correspondence with probability bounded away from zero. In the following result, we describe an estimator that correctly recovers the true correspondence above the information-theoretic threshold.

Theorem 3.1. *Fix constants $\alpha, \beta > 0$ and $s \in [0,1]$. Let $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$. Let $\widehat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2 (that is, the number of matched pairs of vertices for which an edge exists between them in both graphs). If*

$$s^2 \left(\frac{\alpha + \beta}{2}\right) > 1, \tag{7}$$

then

$$\lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_*\right) = 1.$$

We remark that the estimator $\widehat{\pi}$ used in Theorem 3.1 is a natural and well-motivated estimator for the latent mapping π_*. It was first considered by Pedarsani and Grossglauser [49] in the context of the correlated Erdős-Rényi model, and it was later shown in the works [14, 15, 45] that it is in fact the maximum a posteriori (MAP) estimate for π_* in the correlated Erdős-Rényi model. As a result, it achieves the information-theoretic threshold for exact recovery of π_* in the correlated Erdős-Rényi model [15, 55]. The estimator has also been studied in the context of correlated SBMs by Onaran, Erkip and Garg [45]; they show that if the commmunity labels of all vertices in G_1 and G_2 are known, then the permutation which maximizes the number of agreeing edges *and* is consistent with the community labels (i.e., does not map a vertex with label $+1$ to a vertex of label -1) succeeds in recovering π_* exactly if the condition (5) holds. Theorem 3.1 improves on this result using a more refined analysis, and does not assume any prior knowledge of community information.

Our next result proves a converse to Theorem 3.1.

Theorem 3.2. *Fix constants $\alpha, \beta > 0$ and $s \in [0,1]$. Let $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$ and suppose that*

$$s^2 \left(\frac{\alpha + \beta}{2}\right) < 1. \tag{8}$$

Then for any estimator $\widetilde{\pi}(G_1, G_2)$, we have that

$$\lim_{n \to \infty} \mathbb{P}(\widetilde{\pi}(G_1, G_2) = \pi_*) = 0.$$

Together, Theorems 3.1 and 3.2 establish the fundamental information-theoretic limits for exact recovery of π_*. It is a natural generalization of known results for correlated Erdős-Rényi graphs: when $\alpha = \beta$, the same estimator $\widehat{\pi}$ succeeds if $s^2\alpha > 1$, else if $s^2\alpha < 1$ then no estimator recovers π_* with probability bounded away from zero [15, 55].

An overview of the proofs of Theorems 3.1 and 3.2 is given in Section 5; for full proofs see the Supplementary Material.

3.2 Exact Community Recovery

Our first result formalizes the arguments in Section 2 to derive the achievability region of community recovery with two correlated SBMs. The strategy is to first perform exact graph matching, combine the two graphs by overlaying them with respect to π_*, and finally to run a community recovery algorithm on the new graph.

Theorem 3.3. *Fix constants $\alpha, \beta > 0$ and $s \in [0,1]$. Let $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$. Suppose that* (7) *holds and*

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| \geq \sqrt{\frac{2}{1 - (1-s)^2}}. \tag{9}$$

Then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that

$$\lim_{n \to \infty} \mathbb{P}\left(O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1\right) = 1.$$

Proof. Given a permutation π mapping $[n]$ to $[n]$, we let $G_1 \vee_\pi G_2$ be the *union graph with respect to* π, so that (i,j) is an edge in $G_1 \vee_\pi G_2$ if and only if (i,j) is edge in G_1 or $(\pi(i), \pi(j))$ is an edge

in G_2. In the special case where $\pi = \pi_*$, $G_1 \vee_\pi G_2$ is the subgraph of the parent graph G consisting of edges that are in either G_1 or G_2'. Denoting $H_* := G_1 \vee_{\pi_*} G_2$, it is readily seen that

$$H_* \sim \mathrm{SBM}\left(n, \alpha(1-(1-s)^2)\frac{\log n}{n}, \beta(1-(1-s)^2)\frac{\log n}{n}\right). \tag{10}$$

The algorithm we study first computes $\widehat{\pi}(G_1, G_2)$ according to Theorem 3.1. We then pick any community recovery algorithm that is known to succeed until the information-theoretic limit, and run it on $\widehat{H} := G_1 \vee_{\widehat{\pi}} G_2$; we denote the result of this algorithm by $\widehat{\boldsymbol{\sigma}}(\widehat{H})$. We can then write

$$\begin{aligned}\mathbb{P}(O(\widehat{\boldsymbol{\sigma}}(\widehat{H}), \boldsymbol{\sigma}) \neq 1) &\leq \mathbb{P}(\{O(\widehat{\boldsymbol{\sigma}}(\widehat{H}), \boldsymbol{\sigma}) \neq 1\} \cap \{\widehat{H} = H_*\}) + \mathbb{P}(\widehat{H} \neq H_*) \\ &\leq \mathbb{P}(O(\widehat{\boldsymbol{\sigma}}(H_*), \boldsymbol{\sigma}) \neq 1) + \mathbb{P}(\widehat{\pi} \neq \pi_\star),\end{aligned}$$

where, to obtain the inequality in the second line, we have used that $\widehat{\boldsymbol{\sigma}}(\widehat{H}) = \widehat{\boldsymbol{\sigma}}(H_*)$ on the event $\{\widehat{H} = H_*\}$, and that $\widehat{H} \neq H_*$ implies $\widehat{\pi} \neq \pi_*$. Since (9) is the precise condition for when exact community recovery on H_* is possible, we know that $\mathbb{P}(O(\widehat{\boldsymbol{\sigma}}(H_*), \boldsymbol{\sigma}) \neq 1) \to 0$ as $n \to \infty$. Since we also have $\mathbb{P}(\widehat{\pi} \neq \pi_*) \to 0$ in light of (7), this concludes the proof. □

In the regime $s^2(\alpha+\beta)/2 > 1$, Theorem 3.1 proves the existence of an algorithm that can recover communities using *both* G_1 and G_2 when it is information-theoretically impossible to do so using G_1 or G_2 alone. To complement the achievability result of Theorem 3.3, our next result provides conditions under which community recovery is information-theoretically impossible.

Theorem 3.4. *Fix constants $\alpha, \beta > 0$ and $s \in [0,1]$. Let $(G_1, G_2) \sim \mathrm{CSBM}\left(n, \alpha\frac{\log n}{n}, \beta\frac{\log n}{n}, s\right)$ and suppose that*

$$\left|\sqrt{\alpha} - \sqrt{\beta}\right| < \sqrt{\frac{2}{1-(1-s)^2}}. \tag{11}$$

Then for any estimator $\widetilde{\boldsymbol{\sigma}} = \widetilde{\boldsymbol{\sigma}}(G_1, G_2)$, we have that $\lim_{n\to\infty} \mathbb{P}(O(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 0$.

The idea behind the proof is that given

$$H \sim \mathrm{SBM}\left(n, \alpha(1-(1-s)^2)\frac{\log n}{n}, \beta(1-(1-s)^2)\frac{\log n}{n}\right),$$

one can subsample the edges of H to obtain H_1, H_2'. We then generate another graph H_2 by applying a random permutation to the vertex set of H_2'. This can be done so that (H_1, H_2) has the same distribution as (G_1, G_2). If there exists an algorithm for recovering $\boldsymbol{\sigma}$ from (H_1, H_2), this would imply that $\boldsymbol{\sigma}$ can be recovered from H, since (H_1, H_2) are derived from H. However, this is known to be impossible under the condition (11) (see [1]). See the Supplementary Material for the full proof.

We remark that Theorem 3.4 provides a *partial converse* to the achievability result in Theorem 3.3: it is tight when $s^2(\alpha+\beta)/2 > 1$, but the precise information-theoretic threshold is unknown when $s^2(\alpha+\beta)/2 \leq 1$, which is the regime where exact graph matching fails. This leads to an interesting follow-up question: is exact graph matching *necessary* for the exact recovery of communities? We conjecture that it is not, which is formalized as follows.

Conjecture 3.5. *There exists $\epsilon = \epsilon(\alpha, \beta, s) > 0$ such that if* (9) *holds and*

$$s^2\left(\frac{\alpha+\beta}{2}\right) \geq 1 - \epsilon, \tag{12}$$

then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that $\lim_{n\to\infty} \mathbb{P}(O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$.

In words, we believe that the communities can be exactly recovered even in regimes where exact graph matching is information-theoretically impossible. We outline a possible way to prove the conjecture. The algorithm we shall use is the same one used in the proof of Theorem 3.3: we compute $\widehat{\pi}$, the permutation which maximizes the number of agreeing edges across G_1 and G_2, and then run an optimal community recovery algorithm on the union graph $\widehat{H} = G_1 \vee_{\widehat{\pi}} G_2$. Define the *correctly-matched region* $\mathcal{C} := \{i \in [n] : \widehat{\pi}(i) = \pi_*(i)\}$. When $s^2(\alpha+\beta)/2 < 1$, we have that $\mathcal{C} \neq [n]$ with high probability. However, we expect that $|\mathcal{C}| = (1-o(1))n$; that is, $\widehat{\pi}$ coincides with

π_* on all but a negligible fraction of vertices. This is the case in correlated Erdős-Rényi graphs [55], so we expect it to hold for correlated SBMs as well. Let $\widehat{H}_{\mathcal{C}}$ be the subgraph of $\widehat{H}$ restricted to the vertices in $\mathcal{C}$. Since all vertices in $\mathcal{C}$ have been correctly matched, we expect that (possibly in an *approximate* sense)

$$\widehat{H}_{\mathcal{C}} \sim \mathrm{SBM}\left(|\mathcal{C}|, \alpha(1-(1-s)^2)\frac{\log n}{n}, \beta(1-(1-s)^2)\frac{\log n}{n}\right). \tag{13}$$

In particular, if (9) holds, the communities of vertices in $\mathcal{C}$ can be *exactly* recovered. For vertices not in $\mathcal{C}$, note that most of the neighbors will be elements of $\mathcal{C}$, which will have correct community labels. If $\alpha \geq \beta$, the true community label of a given vertex is the same as the true label of most neighbors with high probability (when $\alpha < \beta$, the reverse is true) [2], hence the community memberships of vertices not in $\mathcal{C}$ will be correctly identified as well.

Making the arguments above formal is a challenging task. For one, though we may expect (13) to hold if $\mathcal{C}$ is a *fixed* set, it is in fact a *random* set depending on G_1, G_2, π_* so formally proving (13) will require a careful analysis. Moreover, we would like to use (13) to argue that running a community recovery algorithm on $\widehat{H}$ (rather than $\widehat{H}_C$) perfectly recovers the communities in $\mathcal{C}$. Rigorously justifying these points requires significant effort, so we leave it to future work.

3.3 Multiple correlated stochastic block models

In this section, we outline how one can recover communities using K correlated stochastic block models, again using graph matching as a subroutine.

Formally, we generate $(G_1, \ldots, G_K) \sim \mathrm{CSBM}(n, p, q, s, K)$ via the following construction. First, for $2 \leq i \leq K$, generate i.i.d. uniform random permutations on $[n]$, given by $\pi_*^2, \ldots, \pi_*^K$. We independently generate a parent graph $G \sim \mathrm{SBM}(n, p, q)$. Now, we construct G_1 as well as $G_2', \ldots, G_K'$ independently conditioned on G by including each edge in G with probability s. Finally, for $2 \leq i \leq K$, we generate G_i from G_i' by permuting the vertex labels according to π_*^i.

As in the case of two correlated graphs, the achievability and impossibility results depend on the structure of the union graph with respect to the true permutations $\pi_*^2, \ldots, \pi_*^K$.

Theorem 3.6. *Let* $(G_1, \ldots, G_K) \sim \mathrm{CSBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, s\right)$. *Suppose that* (7) *holds and*

$$|\sqrt{\alpha} - \sqrt{\beta}| \geq \sqrt{\frac{2}{1-(1-s)^K}}. \tag{14}$$

Then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, \ldots, G_K)$ *such that* $\lim_{n\to\infty} \mathbb{P}(O(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$.

As in the case of two graphs, Theorem 3.6 shows that in the regime $s^2(\alpha+\beta)/2 > 1$, exact recovery of communities is possible by combining information from the K networks in cases where it is information-theoretically impossible to do so using a single graph.

Our next result establishes an impossibility result which is analogous to Theorem 3.4.

Theorem 3.7. *Let* $(G_1, \ldots, G_K) \sim \mathrm{CSBM}\left(n, \alpha\frac{\log n}{n}, \beta\frac{\log n}{n}, s\right)$ *and suppose that*

$$|\sqrt{\alpha} - \sqrt{\beta}| < \sqrt{\frac{2}{1-(1-s)^K}}. \tag{15}$$

Then for any estimator $\widetilde{\boldsymbol{\sigma}} = \widetilde{\boldsymbol{\sigma}}(G_1, G_2)$, *we have that* $\lim_{n\to\infty} \mathbb{P}(O(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 0$.

The proofs of Theorems 3.6 and 3.7 are in the Supplementary Material.

We highlight a few interesting aspects of Theorems 3.6 and 3.7. As in the two-graph case, Theorem 3.7 provides a *partial* converse to the achievability result in Theorem 3.6: it is tight in the regime $s^2(\alpha+\beta)/2 > 1$, but the correct threshold remains unknown when $s^2(\alpha+\beta)/2 \leq 1$. Additionally, as K increases, the achievability and impossibility conditions in (14) and (15) converge to the conditions $|\sqrt{\alpha} - \sqrt{\beta}| \geq \sqrt{2}$ and $|\sqrt{\alpha} - \sqrt{\beta}| < \sqrt{2}$, which are the information-theoretic conditions for achievability and impossibility of community recovery in the *parent graph*. In words, the more correlated graphs we observe, the less information is lost when generating the observed graphs from the parent graph via the sampling process.

4 Related work

Our work naturally draws upon techniques in the community recovery literature as well as the graph matching literature. Here, we elaborate on relevant work in these fields that were not covered during the exposition of our model and main results.

Graph Matching. Most of the theoretical literature on graph matching has focused on correlated Erdős-Rényi graphs, which was first introduced by Pedarsani and Grossglauser [49]. Significant progress has been made in recent years in characterizing the fundamental information-theoretic limits for recovering the latent vertex correspondence π_*. Cullina and Kiyavash [15, 14] first derived the precise information-theoretic conditions for *exact* recovery of π_* for sparse graphs (in a sublinear-degree regime) and later Wu, Xu, and Yu [55] refined this to include linear degree regimes. We remark that weaker notions of recovery (e.g., almost exact recovery, partial recovery) have also been addressed; see [16, 24, 26, 25, 55] for more details.

A major open question is whether there exist *efficient algorithms* for computing or approximating π_* in correlated Erdős-Rényi graphs. In particular, the estimators which are known to succeed up to the information-theoretic threshold are usually given by the solution to a combinatorial optimization problem, for which a brute force search takes $O(n!)$ time. Significant improvements were recently made by [43, 6], who provided $n^{O(\log n)}$ time algorithms for exactly recovering π_*. For values of s close to 1, recent work provides polynomial-time algorithms for exact recovery [20, 22, 35].

Community recovery in Multi-layer SBMs. We briefly review the literature on multi-layer SBMs as it is a form of side information studied in the literature that is closest to our work. In this model, first proposed by Han, Xu, and Airoldi [27], a community labeling $\boldsymbol{\sigma}$ is chosen at random. A collection of conditionally independent SBMs on the same vertex set with the same latent community labels are then generated, possibly with different (but known) edge formation probabilities. Variants of this model have been explored by several authors; see, e.g., [5, 47, 48, 32, 4, 7]. The works [38, 34] additionally consider node-level information that is correlated with the latent community membership. While our work also considers multiple networks as side information, we emphasize that there are significant differences. For one, the networks we consider are not conditionally independent given the latent communities, but are also correlated through the formation of edges. Moreover, the node labels are known in the multi-layer setting which completely removes the need for graph matching.

5 Overview of graph matching proofs

5.1 Achievability of exact graph matching: Proof sketch of Theorem 3.1

Let $\mathcal{F}_\epsilon := \{(1-\epsilon)n/2 \le |V_+|, |V_-| \le (1+\epsilon)n/2\}$ denote the event that the two communities are approximately balanced. Since the community labels are i.i.d. uniform, we have for any fixed $\epsilon > 0$ that $\mathbb{P}(\mathcal{F}_\epsilon) = 1 - o(1)$ as $n \to \infty$; we may thus condition on $\mathcal{F}_\epsilon$. Let S_{k_1,k_2} be the set of permutations which mismatches k_1 vertices in V_+ and k_2 vertices in V_-. If (7) holds, there exists $\epsilon = \epsilon(\alpha, \beta, s)$ sufficiently small so that

$$\mathbb{P}(\widehat{\pi} \in S_{k_1,k_2} \,|\, \mathcal{F}_\epsilon) \le n^{-\epsilon(k_1+k_2)}. \tag{16}$$

To bound the probability that $\widehat{\pi} \neq \pi_*$, we then take a union bound over all the events $\{\widehat{\pi} \in S_{k_1,k_2}\}$ such that $k_1 + k_2 \ge 1$, that is, there is at least one mismatched vertex, concluding the proof.

The key technical result which enables the proof is (16); this is derived by deriving tight bounds for the generating function corresponding to the number of agreeing edges in G_1 and G_2 with respect to a given permutation. In prior work on the graph matching problem in correlated Erdős-Rényi graphs, the aforementioned generating functions could be exactly computed [14, 15, 55]. An important difference between work on those models and ours is that the stochastic block model is *heterogeneous*: the probability of edge formation is not i.i.d. over all vertex pairs, but can vary depending on the latent community memberships of the vertex pairs. To handle this heterogeneity, we develop new techniques for bounding the generating functions of interest. Specifically, we derived recursive bounds for the generating functions of interest as a function of the number of vertices, n; see the Supplementary Material for details. We conjecture that this method can be extended to analyze other classes of correlated networks with heterogeneous structure.

5.2 Impossibility of exact graph matching: Proof sketch of Theorem 3.2

Let H be the intersection graph between G_1 and G_2', that is, (i, j) is an edge in H if and only if (i, j) is an edge in G_1 and G_2'. Equivalently, (i, j) must be an edge in the parent graph G and must be included in both G_1 and G_2'. Since the probability of the latter event is s^2, we see that

$$H \sim \mathrm{SBM}\left(n, \alpha s^2 \frac{\log n}{n}, \beta s^2 \frac{\log n}{n}\right).$$

If $s^2(\alpha + \beta)/2 < 1$, then H is not connected with probability tending to 1 as $n \to \infty$. In particular, we can show that there are many singletons in H, which are vertices that have *non-overlapping neighborhoods* in G_1 and G_2'. Due to the lack of shared information, it is difficult to match such vertices across the two graphs, even for optimal estimators that have access to the ground-truth community labeling $\boldsymbol{\sigma}$. To be precise, we study the performance of the maximum a posteriori (MAP) estimator of π_* given G_1, G_2, *and* $\boldsymbol{\sigma}$, which minimizes the probability of error. Since the MAP estimator cannot output π_* with probability bounded away from zero, neither can *any* other estimator.

We remark that the strategy of studying the connectivity threshold of the intersection graph was previously employed by Cullina and Kiyavash for correlated Erdős-Rényi graphs [14, 15]. We expect that this strategy can be used to prove converse arguments in other models of correlated random graphs (e.g., the Chung-Lu model) as well. We leave this for future work.

6 Discussion and future work

In this work, we studied the problem of exact community recovery given multiple correlated SBMs as side information. Specifically, our goal was to understand how this side information changes the fundamental information-theoretic threshold for achievability and impossibility of exact community recovery. Strikingly, using multiple correlated SBMs allows one to exactly recover communities in regimes where it is information-theoretically impossible to do so using a single graph.

Precisely, we determine the sharp information-theoretic condition for exact graph matching in a pair of correlated SBMs. We then apply this to determine conditions for achievability and impossibility of exact community recovery. In the regime where exact graph matching is achievable, we identify the exact information-theoretic conditions for achievability and impossibility of exact community recovery. We also discuss extensions where $K \geq 2$ correlated SBMs are given as side information.

Our work leaves open several important avenues for future work, which we outline below.

Closing the information-theoretic gaps in exact community recovery. Together, Theorems 3.3 and 3.4 show that in the regime $s^2(\alpha + \beta)/2 > 1$, we have identified the information-theoretic threshold between impossibility and achievability. However, we do not have achievability results for the regime $s^2(\alpha + \beta)/2 < 1$, since exact graph matching is not possible in this case. This leads to the following natural question which is formalized in Conjecture 3.5: *is exact graph matching needed for exact community recovery?* We believe the answer is *no*; we expect that showing this rigorously will lead to new algorithms for jointly synthesizing networks and identifying communities.

Efficient algorithms. Our achievability algorithms rely on graph matching as a subroutine, which is computationally expensive. Do there exist *efficient* algorithms for graph matching in the correlated SBM model? If not, is it possible to recover communities exactly using a *polynomial-time relaxation* of the graph matching subroutine?

General correlated stochastic block models. For simplicity of exposition, we focused on the simplest setting of the stochastic block model where there are two balanced communities. A natural future direction is to extend our results to account for more general SBMs with multiple communities.

Beyond exact community recovery. Besides exact recovery, other natural notions of community recovery include almost exact recovery, where the goal is to recover all but a negligible fraction of community labels, and partial recovery, where the goal is to recover more than half of the community labels. Using correlated networks as side information to accomplish these tasks is a natural and exciting direction. A key challenge is that in the regimes where phase transitions occur for almost exact and partial recovery, exact graph matching is information-theoretically impossible [1], hence our methods for establishing achievability for community recovery will not work. Solving this problem will lead to new methods for community detection based on data from multiple networks.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[3] E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688, 2015.

[4] H. T. Ali, S. Liu, Y. Yilmaz, R. Couillet, I. Rajapakse, and A. Hero. Latent heterogeneous multilayer community detection. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8142–8146, 2019.

[5] J. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. Preprint available at `https://arxiv.org/abs/1906.10026`, 2020.

[6] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng. (Nearly) Efficient Algorithms for the Graph Matching Problem on Correlated Random Graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9190–9198, 2019.

[7] S. Bhattacharyya and S. Chatterjee. Consistent Recovery of Communities from Sparse Multi-relational Networks: A Scalable Algorithm with Optimal Recovery Conditions. In *Complex Networks XI*, pages 92–103, 2020.

[8] R. B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 280–285, 1987.

[9] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1347–1357, 2015.

[10] T. Bui, S. Chaudhuri, T. Leighton, and M. Sipser. Graph Bisection Algorithms With Good Average Case Behavior. In *25th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 181–192, 1984.

[11] J. Chen and B. Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[12] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.

[13] T. Cour, P. Srinivasan, and J. Shi. Balanced graph matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 313–320, 2007.

[14] D. Cullina and N. Kiyavash. Improved Achievability and Converse Bounds for Erdős-Rényi Graph Matching. In *ACM SIGMETRICS*, volume 44, pages 63–72, 2016.

[15] D. Cullina and N. Kiyavash. Exact alignment recovery for correlated Erdős-Rényi graphs. Preprint available at `https://arxiv.org/abs/1711.06783`, 2018.

[16] D. Cullina, N. Kiyavash, P. Mittal, and H. V. Poor. Partial Recovery of Erdős-Rényi Graph Alignment via k-Core Alignment. *SIGMETRICS Perform. Eval. Rev.*, 48(1):99–100, July 2020.

[17] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[18] Y. Deshpande, S. Sen, A. Montanari, and E. Mossel. Contextual Stochastic Block Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8581–8593, 2018.

[19] I. S. Dhillon, Y. Guan, and B. J. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

[20] J. Ding, Z. Ma, Y. Wu, and J. Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115, 2021.

[21] M. Dyer and A. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

[22] Z. Fan, C. Mao, Y. Wu, and J. Xu. Spectral graph matching and regularized quadratic relaxations: Algorithm and theory. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 2985–2995. PMLR, 13–18 Jul 2020.

[23] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[24] L. Ganassali and L. Massoulié. From tree matching to sparse graph alignment. In *Proceedings of the Thirty Third Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1633–1665. PMLR, 09–12 Jul 2020.

[25] L. Ganassali, L. Massoulié, and M. Lelarge. Impossibility of Partial Recovery in the Graph Alignment Problem. Preprint available at `https://arxiv.org/abs/2102.02685`, 2021.

[26] G. Hall and L. Massoulié. Partial Recovery in the Graph Alignment Problem. Preprint available at `https://arxiv.org/abs/2007.00533`, 2020.

[27] Q. Han, K. Xu, and E. Airoldi. Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning (ICML)*, pages 1511–1520. PMLR, 2015.

[28] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[29] V. Kanade, E. Mossel, and T. Schramm. Global and Local Information in Clustering Labeled Block Models. *IEEE Transactions on Information Theory*, 62(10):5906–5917, 2016.

[30] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[31] N. Korula and S. Lattanzi. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*, 7(5):377–388, 2014.

[32] J. Lei, K. Chen, and B. Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 12 2019.

[33] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[34] Z. Ma and S. Nandy. Community Detection with Contextual Multilayer Networks. Preprint available at `https://arxiv.org/abs/2104.02960`, 2021.

[35] C. Mao, M. Rudelson, and K. Tikhomirov. Random Graph Matching with Improved Noise Robustness. Preprint available at `https://arxiv.org/abs/2101.11783`, 2021.

[36] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[37] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 694–703. ACM, 2014.

[38] V. Mayya and G. Reeves. Mutual information in community detection with covariate information and correlated networks. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 602–607, 2019.

[39] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162, 07 2014.

[40] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. *Electronic Journal of Probability*, 21(none):1 – 24, 2016.

[41] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[42] E. Mossel and J. Xu. Local Algorithms for Block Models with Side Information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 71–80, 2016.

[43] E. Mossel and J. Xu. Seeded graph matching via large neighborhood statistics. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1005–1014, 2019.

[44] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE Computer Society, 2009.

[45] E. Onaran, S. Garg, and E. Erkip. Optimal de-anonymization in random graphs with community structure. In *2016 IEEE 37th Sarnoff Symposium*, pages 1–2, 2016.

[46] E. Onaran, S. Garg, and E. Erkip. Optimal de-anonymization in random graphs with community structure. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 709–713. IEEE, 2016.

[47] S. Paul and Y. Chen. Null Models and Community Detection in Multi-Layer Networks. Preprint available at `https://arxiv.org/abs/1608.00623`, 2020.

[48] S. Paul and Y. Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230 – 250, 2020.

[49] P. Pedarsani and M. Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1235–1243, 2011.

[50] Y. Ruan, D. Fuhry, and S. Parthasarathy. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 1089–1098, New York, NY, USA, 2013.

[51] H. Saad and A. Nosratinia. Recovering a single community with side information. *IEEE Transactions on Information Theory*, 66(12):7939–7966, 2020.

[52] S. Sahebi and W. Cohen. Community-based recommendations: a solution to the cold start problem. In *Workshop on Recommender Systems and the Social Web (RSWEB), held in conjunction with ACM RecSys'11*, October 2011.

[53] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[54] R. Wu, J. Xu, R. Srikant, L. Massoulie, M. Lelarge, and B. Hajek. Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '15, page 449–450, 2015.

[55] Y. Wu, J. Xu, and S. H. Yu. Settling the Sharp Reconstruction Thresholds of Random Graph Matching. Preprint available at `https://arxiv.org/abs/2102.00082`, 2021.

[56] L. Yu, J. Xu, and X. Lin. The Power of D-hops in Matching Power-Law Graphs. Preprint available at `https://arxiv.org/abs/2102.12975`, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See page 4.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The main text provides a full proof of Theorem 3.3 and proof sketches of the other theorems. Full proofs are in the Supplementary Material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Relevant prior work on models (all theoretical) is cited and discussed in detail throughout the paper.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]