Achievability Bounds for Community Detection and Matrix Completion with Two-Sided Graph Side-Information

Qiaosheng Zhang
Department of ECE
National University of Singapore
Email: elezqiao@nus.edu.sg

Vincent Y. F. Tan
Department of ECE/Mathematics
National University of Singapore
Email: vtan@nus.edu.sg

Changho Suh
School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Email: chsuh@kaist.ac.kr

***Abstract*—We consider the problem of recovering communities of users and communities of items (such as movies) based on a partially observed rating matrix as well as side-information in the form of similarity graphs of the users and items. The user-to-user and item-to-item similarity graphs are generated according to the celebrated stochastic block model (SBM). We develop a lower bound on the minimum expected number of observed ratings (also known as the *sample complexity*) needed for this recovery task, which is a function of various parameters including the quality of the graph side-information manifested in the intra- and inter-cluster probabilities of the SBMs. Our information-theoretic results quantify the benefits of the two-sided graph side-information for recovery, and further analysis reveals that the two pieces of graph side-information produce an interesting synergistic effect under certain scenarios. This means that if one observes only *one* of the two graphs, then the required sample complexity worsens to the case in which *none* of the graphs is observed. Thus *both* graphs are strictly needed to reduce the sample complexity.**

***Index Terms*—Community detection, matrix completion, stochastic block model, graph side-information.**

Full version [1]: https://arxiv.org/pdf/1912.04099.pdf

I. Introduction

Recommender systems aim to accurately predict users' preferences and recommend appropriate items for users based on available data that is usually scant and/or of low quality. For example, Nexflix's movie recommender system relies heavily on the *rating matrix* that comprises users' evaluations of movies, and various recommendation algorithms (such as collaborative filtering [2]) have been developed. However, merely exploiting the available ratings may not be sufficient for high-quality recommendations, since (i) the rating matrices are usually highly incomplete, and (ii) the preferences of new users are always unavailable (i.e., the *cold start problem*). Meanwhile, it has been noticed that community information—either the *communities of users* (e.g., the friendships in Facebook) or the *communities of movies* (e.g., the categories/genres of movies in the Netflix database)—may effectively improve the quality of recommendations [3], [4] and tackle the cold start problem [5], as users in the same community are more likely to share similar preferences, and movies in the same community are more likely to attract similar users.

This work of Vincent Y. F. Tan is funded by a Singapore National Research Foundation (NRF) Fellowship (R-263-000-D02-281). This work of Changho Suh is supported in part by the Air Force Office of Scientific Research under award number FA2386-19-1-4050.

While most of the attention has focused on the algorithmic developments of the graph-aided recommender systems as well as their accompanying analyses, the fundamental limits of such problems are also worth exploring. Ahn et al. [4] considered the problem of recovering the binary rating matrix (which comprises users' ratings to movies) based on a partially observed matrix and a user-to-user similarity graph. They characterized a sharp threshold on the sample complexity needed for recovery as a function of the quality of the graph and the amount of noise in the measurements, and also quantified the gains due to graph side-information. In practice, the item-to-item similarity graph can also be constructed from the features of items [6], [7]; hence one may ask whether the additional item-to-item graph provides strictly more benefits, and whether observing *two* pieces of graph side-information simultaneously has synergistic effects. This work precisely addresses these questions by investigating the benefits of the two-sided graph side-information for a recovery problem. To the best of our knowledge, there were no prior works studying the benefits of exploiting two graphs in the graph-aided recommender systems from an information-theoretic perspective.

We consider a concrete example of movie recommender systems with n users and m movies, wherein users' ratings to movies are either 0 (dislike) or 1 (like). Users are partitioned into communities of men and women (of equal size), while movies are partitioned into communities of action movies and romance movies (of equal size). The assumptions on binary ratings and two equal-sized communities are mainly for ease of presentation, and extensions to general settings are certainly also possible. Typically, action movies attract more men and romance movies attract more women, but we also allow the existence of *atypical* action movies and romance movies. The *nominal ratings* from a certain community of users to a certain type of movies are pre-specified in Table I.[1] A *personalized rating* of each individual rating is a perturbed version of the corresponding nominal rating (being flipped with probability less than $1/2$), modeling the different preferences of users in the same community to a certain movie. The $n \times m$ binary

[1]As an initial effort, we assume there is only one type of atypical movies that is completely different from typical ones (e.g., typical action movies attract men while atypical action movies attract women). For future work, one may extend to a more general setting wherein three types of atypical movies with different likabilities are considered (e.g., atypical action movies may attract women only, both men and women, or neither).

	Action movies		Romance movies	
	Typical	Atypical	Typical	Atypical
Men	1	0	0	1
Women	0	1	1	0

TABLE I: Nominal ratings from a certain community of users to a certain type of movies

rating matrix comprises n users' ratings to all the m movies.

Under this setting, three pieces of information are observed by the learner: (i) Entries in the rating matrix that are observed/sampled (independently) with a fixed *sample probability*, (ii) The user-to-user similarity graph that is generated according to the celebrated symmetric *stochastic block model* (SBM) [8], [9] with two equal-sized communities (also known as the *planted bisection model*), and (iii) The item-to-item similarity graph that is generated according to another independent symmetric SBM with two equal-sized communities. The task here is to exactly recover the communities of men and women, the communities of action and romance movies, as well as to uncover the atypical action and romance movies.

Main Contributions: We develop a lower bound on the sample complexity needed for recovery as a function of the qualities of the user-to-user similarity graph (i.e., the row graph) and the item-to-item similarity graph (i.e., the column graph). The qualities of the graphs can be quantified by the difference between the intra- and inter-cluster connection probabilities of the SBMs that govern them. Our theoretical studies show that, from the viewpoint of the sample complexity, gains due to the two-side graph side-information appear for a wide range of parameters. More interestingly, we show that there exists a certain regime in which there is a synergistic effect of the two-sided graph side-information—simultaneously observing *both* graphs is helpful for recovery, while observing *only one* graph is equivalent to observing neither.

Related Works: The *community detection* problem, which aims to partition vertices into different communities (or clusters) based on the density of connections, has been well-studied from an information-theoretic perspective [9]–[14]. Moreover, it has been shown that side-information (e.g., node values [15]–[19], edge weights [20], similarity between data points [21]) is also helpful in recovering hidden communities. In our problem, we aim to recover the communities of users and movies given realizations from two symmetric SBMs with a partially observed matrix. Also, we note that the task in [22] (joint recovery of rows and columns communities) is similar to ours, but therein, the graph information is not available.

The *matrix completion* problem focuses on the recovery of low-rank matrices from sparse observations, and has wide applications in recommender systems [23]. Unlike the standard setting in which the linear dependence of rows and columns of the low-rank matrix is unstructured, the graph side-information in recommender systems also imposes additional structures on the low-rank matrix to be completed. For instance, [4], [24], [25] considered a specific binary matrix completion problem with the aid of one-sided graph side-information (the user-to-user graph), while [6] and [26] studied matrix completion models whereby additional proximity information about both rows and columns is available. The task in this work is strictly more challenging than merely recovering a low-rank matrix (as discussed in Remark 1 in Section II-D); nonetheless, one can view the problem of recovering a low-rank matrix with two-sided graph side-information as a by-product of our task.

II. Problem statement

A. Notation

For any integer $a \geq 1$, $[a]$ represents the set of integers $\{1, \ldots, a\}$. For any integers a, b such that $a < b$, $[a : b]$ represents the set of integers $\{a, a+1, \ldots, b\}$. For any event $\mathcal{E}$, the conditional probability $\mathbb{P}(\cdot|\mathcal{E})$ is abbreviated as $\mathbb{P}_{\mathcal{E}}(\cdot)$.

B. Model

Consider n users and m movies, and we require $m = \omega(\log n)$ and $n = \omega(\log m)$ for technical reasons (which essentially excludes very "fat" and very "thin" matrices). The sets of men and women are respectively denoted by $\mathcal{M}$ and $\mathcal{W}$, where $\mathcal{M}, \mathcal{W} \subset [n]$, $|\mathcal{M}| = |\mathcal{W}| = n/2$, and $\mathcal{M} \cap \mathcal{W} = \emptyset$. The sets of action and romance movies are respectively denoted by $\mathcal{A}$ and $\mathcal{R}$, where $\mathcal{A}, \mathcal{R} \subset [m]$, $|\mathcal{A}| = |\mathcal{R}| = m/2$, and $\mathcal{A} \cap \mathcal{R} = \emptyset$. Meanwhile, there may also exist an unknown-sized subset of atypical action movies $\mathcal{A}_0 \subseteq \mathcal{A}$ and an unknown-sized subset of atypical romance movies $\mathcal{R}_0 \subseteq \mathcal{R}$. Let $\bar{\mathcal{A}} \triangleq \mathcal{A} \setminus \mathcal{A}_0$ and $\bar{\mathcal{R}} \triangleq \mathcal{R} \setminus \mathcal{R}_0$.

Let $\xi_{\mathcal{M},\mathcal{W},\mathcal{A},\mathcal{R},\mathcal{A}_0,\mathcal{R}_0}$ be an aggregation of the parameters of interest, and we sometimes abbreviate it as ξ for notational convenience. The sets of men, women, action and romance movies, atypical action and romance movies (associated with ξ) are respectively denoted by $\xi_{\mathcal{M}}, \xi_{\mathcal{W}}, \xi_{\mathcal{A}}, \xi_{\mathcal{R}}, \xi_{\mathcal{A}_0}, \xi_{\mathcal{R}_0}$. In order to avoid any indeterminacies, without loss of generality,[2] we assume that the majority of the first $n/2$ users are men (i.e., $|\xi_{\mathcal{M}} \cap [n/2]| > n/4$), and the majority of the first $m/2$ movies are action movies (i.e., $|\xi_{\mathcal{A}} \cap [m/2]| > m/4$). The parameter space Ξ is the collection of valid parameters $\xi_{\mathcal{M},\mathcal{W},\mathcal{A},\mathcal{R},\mathcal{A}_0,\mathcal{R}_0}$.

The nominal ratings from users to movies, as stated in Table I, reflects our assumption that typical action movies $\bar{\mathcal{A}}$ and atypical romance movies $\mathcal{R}_0$ attract more men, while typical romance movies $\bar{\mathcal{R}}$ and atypical action movies $\mathcal{A}_0$ attract more women. The taste of each individual user also differs from the nominal taste of the communities. For each individual man or woman, the ratings to action movies are independently perturbed (i.e., flipped from nominal ratings) by Bernoulli random variables Bern($\theta_{\mathcal{A}}$), while the ratings to romance movies are independently perturbed by Bernoulli random variables Bern($\theta_{\mathcal{R}}$), where $\theta_{\mathcal{A}}, \theta_{\mathcal{R}} \in (0, \frac{1}{2})$ are the personalization parameters for action and romance movies, respectively. The difference between $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{R}}$ is an important statistic for distinguishing action and romance movies.

For any $\xi \in \Xi$, the corresponding $n \times m$ *non-personalized binary rating matrix* B_ξ denotes users' nominal ratings to all the movies. Note that B_ξ is uniquely determined by ξ according to Table I (e.g., $(B_\xi)_{ij} = 1$ if $i \in \xi_{\mathcal{M}}, j \in \xi_{\bar{\mathcal{A}}}$, and $(B_\xi)_{ij} = 0$ if $i \in \xi_{\mathcal{W}}, j \in \xi_{\mathcal{R}_0}$), but a same matrix B may correspond to different distinct instances in Ξ (see Remark 1 below for an example). The $n \times m$ *personalized*

[2] Without this assumption, for any ξ, one can always find a ξ' with $\xi'_{\mathcal{W}} = \xi_{\mathcal{M}}$, $\xi'_{\mathcal{R}} = \xi_{\mathcal{A}}$, $\xi'_{\mathcal{R}_0} = \xi_{\mathcal{A}_0}$, and $\xi'_{\mathcal{A}_0} = \xi_{\mathcal{R}_0}$ (i.e., simultaneously flipping the communities of users and movies) such that ξ and ξ' are indistinguishable.

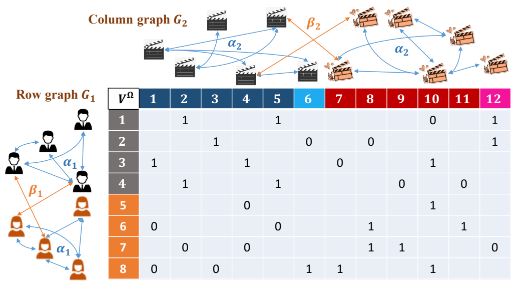


Fig. 1: An illustration of V^Ω, G_1, and G_2 that are generated according to the model parameterized by ξ, where $\xi_{\mathcal{M}} = \{1,2,3,4\}$ (gray), $\xi_{\mathcal{W}} = \{5,6,7,8\}$ (orange), $\xi_{\mathcal{A}} = \{1,2,3,4,5,6\}$ (blue), $\xi_{\mathcal{A}_0} = \{6\}$ (light blue), $\xi_{\mathcal{R}} = \{7,8,9,10,11,12\}$ (red), and $\xi_{\mathcal{R}_0} = \{12\}$ (pink).

binary rating matrix V_ξ denotes all the users' *actual* ratings to all the movies, where $(V_\xi)_{ij}$ equals $(B_\xi)_{ij} \oplus \text{Bern}(\theta_{\mathcal{A}})$ if $j \in \mathcal{A}$, and $(B_\xi)_{ij} \oplus \text{Bern}(\theta_{\mathcal{R}})$ if $j \in \mathcal{R}$. The i-th row of V_ξ is the i-th user's ratings to all the movies, whereas the j-th column of V_ξ is all the users' ratings to the j-th movie.

C. Observations

The learner observes three pieces of information.
(a) The partially observed matrix V^Ω. For each $(i,j) \in [n] \times [m]$, the (i,j)-th entry of V^Ω is given by

$$(V^\Omega)_{ij} = \begin{cases} (V_\xi)_{ij}, & \text{with probability } p, \\ \perp, & \text{with probability } 1-p. \end{cases}$$

where $\perp$ denotes the erasure symbol, and p is the sample probability. The sample complexity then equals nmp, which corresponds to the expected number of observed entries.
(b) The row graph G_1 with n nodes corresponding to the n users. For any pairs of nodes $i \neq j$, they are connected with probability $\alpha_1 = \frac{a_1 \log n}{n}$ if i and j are in the same community, and with probability $\beta_1 = \frac{b_1 \log n}{n}$ otherwise.
(c) The column graph G_2 with m nodes corresponding to the m movies. For any pairs of nodes $i \neq j$, they are connected with probability $\alpha_2 = \frac{a_2 \log m}{m}$ if i and j are in the same community, and with probability $\beta_2 = \frac{b_2 \log m}{m}$ otherwise.

An example of the three pieces of information V^Ω, G_1, and G_2 is illustrated in Fig. 1. Note that a_1, a_2, b_1, b_2 are constants, while $\alpha_1, \beta_1 = \Theta(\frac{\log n}{n})$ and $\alpha_2, \beta_2 = \Theta(\frac{\log m}{m})$. We define $I_1 \triangleq (\sqrt{a_1} - \sqrt{b_1})^2$ as the quality of the row graph G_1, since a larger difference between a_1 and b_1 makes recovery easier. A well-known result [9] shows that exact recovery is possible if $I_1 > 2$ and impossible if $I_1 < 2$. Similarly, we define $I_2 \triangleq (\sqrt{a_2} - \sqrt{b_2})^2$ as the quality of the column graph G_2.

D. Objective

Based on the observations V^Ω, G_1, and G_2, the learner aims to use an estimator $\phi = \phi(V^\Omega, G_1, G_2)$ to recover ξ, including the sets of users ($\xi_{\mathcal{M}}$ and $\xi_{\mathcal{W}}$), the sets of movies ($\xi_{\mathcal{A}}$ and $\xi_{\mathcal{R}}$), and the sets of atypical movies ($\xi_{\mathcal{A}_0}$ and $\xi_{\mathcal{R}_0}$).

Definition 1 (Exact recovery)**.** *For any estimator ϕ, the maximum probability of error is defined as*

$$P_{\text{err}}(\phi) \triangleq \max_{\xi \in \Xi} \mathbb{P}_\xi(\phi(V^\Omega, G_1, G_2) \neq \xi), \tag{1}$$

where $\mathbb{P}_\xi(\cdot)$ represents the probability of error when V^Ω, G_1, and G_2 are generated according to the model parameterized by ξ. An estimator ϕ satisfies the exact recovery property if $P_{\text{err}}(\phi)$ goes to zero as n tends to infinity.

As a by-product, an estimator ϕ with a vanishing $P_{\text{err}}(\phi)$ is also able to reliably recover the non-personalized binary rating matrix B_ξ with high probability. However, as stated in Remark 1 below, the ability of merely recovering the binary rating matrix B_ξ does not suffice for our task.

Remark 1. *For two different instances $\xi \neq \xi'$, their corresponding non-personalized binary rating matrices B_ξ and $B_{\xi'}$ may be the same. This can be shown via the following example with $m = 6, n = 2$:*

- $\xi_{\mathcal{M}} = \{1\}$, $\xi_{\mathcal{W}} = \{2\}$, $\xi_{\mathcal{A}} = \{1,2,3\}$, $\xi_{\mathcal{R}} = \{4,5,6\}$, $\xi_{\mathcal{A}_0} = \{3\}$, *and* $\xi_{\mathcal{R}_0} = \emptyset$;
- $\xi'_{\mathcal{M}} = \{1\}$, $\xi'_{\mathcal{W}} = \{2\}$, $\xi'_{\mathcal{A}} = \{1,2,4\}$, $\xi'_{\mathcal{R}} = \{3,5,6\}$, $\xi'_{\mathcal{A}_0} = \{4\}$, *and* $\xi'_{\mathcal{R}_0} = \emptyset$.

In both cases,

$$B_\xi = B_{\xi'} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Essentially, movies that attracts men may be either typical action movies or atypical romance movies. This observation also applies to movies that attract women. Because of this flexibility in the classification of movies, the ability of recovering B_ξ does not guarantee reliable recovery of ξ.

III. Main Results

For ease of presentation, the following results are stated in terms of the sample probability p (which is proportional to the sample complexity nmp), and we sometimes use these two notions interchangeably. Theorem 1 below provides a lower bound on the sample probability p, as a function of personalization parameters $\theta_{\mathcal{A}}, \theta_{\mathcal{R}}$, and I_1 and I_2 (the qualities of two graphs). We first define three functions of $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{R}}$:

$$\tau_{\mathcal{AR}} \triangleq 1 - \sqrt{\theta_{\mathcal{A}}\theta_{\mathcal{R}}} - \sqrt{(1-\theta_{\mathcal{A}})(1-\theta_{\mathcal{R}})},$$
$$\nu_{\mathcal{A}} \triangleq 1 - 2\sqrt{\theta_{\mathcal{A}}(1-\theta_{\mathcal{A}})}, \quad \nu_{\mathcal{R}} \triangleq 1 - 2\sqrt{\theta_{\mathcal{R}}(1-\theta_{\mathcal{R}})}.$$

Theorem 1. *(a) Consider the regime in which $\theta_{\mathcal{A}} \neq \theta_{\mathcal{R}}$. For any $\epsilon > 0$, if*

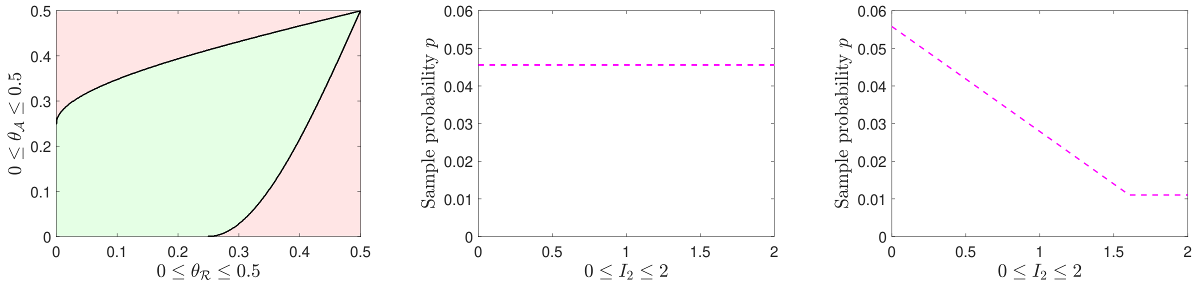
$$p \geq \max\left\{ \frac{(2(1+\epsilon) - I_1)\log n}{(\nu_{\mathcal{A}} + \nu_{\mathcal{R}})m}, \frac{(1+\epsilon)\log m}{\min\{\nu_{\mathcal{A}}, \nu_{\mathcal{R}}\} \cdot n}, \frac{(2(1+\epsilon) - I_2)\log m}{2\tau_{\mathcal{AR}} n} \right\}, \tag{2}$$

there exists a sequence of estimators estimator ϕ satisfying $\lim_{n\to\infty} P_{\text{err}}(\phi) = 0$.

(b) Consider the regime in which $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$. For any $\epsilon > 0$, if $I_2 \geq 2(1+\epsilon)$ and

$$p \geq \max\left\{ \frac{(2(1+\epsilon) - I_1)\log n}{(\nu_{\mathcal{A}} + \nu_{\mathcal{R}})m}, \frac{(1+\epsilon)\log m}{\min\{\nu_{\mathcal{A}}, \nu_{\mathcal{R}}\} \cdot n} \right\}, \tag{3}$$

there exists a sequence of estimators estimator ϕ satisfying $\lim_{n\to\infty} P_{\text{err}}(\phi) = 0$.



(a) Partitions of the $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$-plane. (b) $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.4, 0.1)$ (c) $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.3, 0.15)$

Fig. 2: Fig. 2a shows the dominant term of equation (2) for $I_1 = I_2 = \epsilon = 0$ and different values of $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$: the second and third terms are respectively the dominant term when $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the red and green regions. Fig. 2b and Fig. 2c plot the sample probability p as a function of I_2 (for $n = m = 10,000$ and arbitrary I_1).

In particular, the estimator ϕ in Theorem 1 can be chosen as the *maximum likelihood estimator* ϕ_{ML}, and we sketch the analysis of ϕ_{ML} in Section IV.

Remark 2. *By noting that $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$ implies $\tau_{\mathcal{AR}} = 0$, the achievability result in* (3) *can be interpreted as a limiting consequence of* (2) *as $\theta_{\mathcal{A}} \to \theta_{\mathcal{R}}$. When $I_2 \ge 2(1+\epsilon)$, the third term of* (2) *is non-positive and thus plays no role in the overall expression. When $I_2 < 2$, no achievability result is provided since the third term of* (2) *becomes infinity.*

To distinguish movies that attract same community of users (e.g., typical action movies and atypical romance movies), one wishes to use both the column graph G_2 and the partially observed matrix V^{Ω} governed by $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{R}}$. However, the matrix V^{Ω} becomes useless in distinguishing such movies when $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$, hence the column graph G_2 must be good enough (i.e., $I_2 \ge 2(1+\epsilon)$ as stated in Theorem 1) so that communities of movies can be exactly recovered based on G_2 only.

Example 1 ($n = m$)**.** Note that regardless of the values of I_1 and $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$, in both (2) and (3), the first term $\frac{(2(1+\epsilon)-I_1)\log n}{(\nu_{\mathcal{A}}+\nu_{\mathcal{R}})m}$ is always upper bounded by the second term $\frac{(1+\epsilon)\log m}{\min\{\nu_{\mathcal{A}},\nu_{\mathcal{R}}\}\cdot n}$. This implies that when $n = m$, the first term is inactive, and observing the row graph G_1 (with $I_1 > 0$) does not help to reduce the sample complexity compared to the scenario in which G_1 is not observed. In fact, the above observation can be generalized to every (m, n)-pair such that $n \le m$.

Another natural question to ask is that whether observing the column graph G_2 (with $I_2 > 0$) helps to reduce the sample complexity. We assume the slackness parameter $\epsilon = 0$, and analyze three different cases in the following.

(a) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the red region in Fig. 2a, the right-hand side (RHS) of (2) is dominated by the second term regardless of the value of I_2, hence observing the column graph G_2 does not reduce the sample probability. Fig. 2b plots the sample probability p as a function of I_2 for $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.4, 0.1)$, and note that p stays constant as I_2 increases. This conclusion intuitively makes sense since the "big difference" between $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{R}}$ makes it easy to distinguish the action and romance movies from the partially observed matrix V^{Ω}, and the column graph G_2 then becomes useless.

(b) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the green region in Fig. 2a and satisfies $\theta_{\mathcal{A}} \ne \theta_{\mathcal{R}}$, observing the column graph G_2 does help to reduce the sample probability. This is because the sample probability is dominated by the third term of (2) when $I_2 = 0$, and increasing I_2 effectively decreases the third term. As illustrated in Fig. 2c for $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.3, 0.15)$, the sample probability with any positive I_2 is strictly smaller than that with $I_2 = 0$. Another interesting phenomenon is that once I_2 exceeds the "threshold" $2 - \frac{2\tau_{\mathcal{AR}}}{\min\{\nu_{\mathcal{A}},\nu_{\mathcal{R}}\}}$ (which is strictly positive), the second term of (2) then becomes active, and the gain of increasing I_2 saturates. In Fig. 2c, as I_2 increases, the sample probability first decreases and then stays constant.

(c) When $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$, exact recovery is possible if $I_2 > 2$. Thus, observing the column graph G_2 is helpful only when the quality of G_2 is sufficiently high (i.e., $I_2 > 2$).

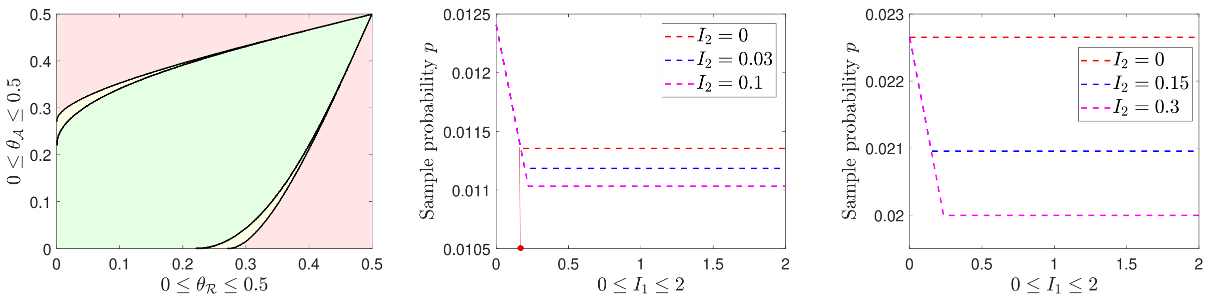
Example 2 ($n = 5m$)**.** Again, we assume the slackness parameter $\epsilon = 0$, and analyze several cases as follows.

(a) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the red region (including the boundary between the red and yellow regions) in Fig. 3a, the RHS of (2) is dominated by the second term regardless of the values of (I_1, I_2), hence neither the row graph G_1 nor the column graph G_2 helps to reduce the sample probability.

(b) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the yellow region in Fig. 3a, the RHS of (2) is dominated by the first term, hence observing the row graph G_1 (with $I_1 > 0$) reduces the sample probability. Fig. 3b plots the sample probability p as a function of I_1 for $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.3, 0.03)$ and three different values of I_2. Note that (i) regardless of the value of I_2, the sample probability with any positive I_1 is strictly smaller than that with $I_1 = 0$; and (ii) the column graph G_2 (with $I_2 > 0$) is also helpful when I_1 exceeds the red point, and is useless otherwise.

(c) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the green region in Fig. 3a and satisfies $\theta_{\mathcal{A}} \ne \theta_{\mathcal{R}}$, the RHS of (2) is dominated by the third term, hence observing the column graph G_2 (with $I_2 > 0$) reduces the sample probability.

(d) When $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ is on the boundary between the yellow and green regions in Fig. 3a, the first and the third terms in (2) are equal and RHS of (2) is dominated by both of the two terms. In this regime, observing *both* the row and column graphs G_1 and G_2 (with $I_1 > 0$, $I_2 > 0$) reduces the sample probability compared to the scenario in which neither is observed. More interestingly, observing *only one* of the two graphs is equivalent to observing neither. Thus, there is a synergistic effect when both pieces of side-information (i.e., both graphs) are



(a) Partitions of the $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$-plane. (b) $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.3, 0.03)$ (c) $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.35, 0.1156)$

Fig. 3: Fig. 3a shows the dominant term of equation (2) for $I_1 = I_2 = \epsilon = 0$ and different values of $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$. The first, the second, and the third terms are respectively the dominant term when $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$ falls into the yellow region, the red region, and the green region. Fig. 3b and Fig. 3c plot the sample probability p as a function of I_1 for $n = 5m = 10,000$ and different values of $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}})$.

observed. The above argument is also illustrated in Fig. 3c for the boundary point $(\theta_{\mathcal{A}}, \theta_{\mathcal{R}}) = (0.35, 0.1156)$.

IV. Proof Sketch of Achievability

The maximum likelihood estimator ϕ_{ML} can be used to reconstruct $\xi_{\mathcal{M},\mathcal{W},\mathcal{A},\mathcal{R},\mathcal{A}_0,\mathcal{R}_0}$, and as long as the sample probability p exceeds the lower bound in equation (2) (for $\theta_{\mathcal{A}} \neq \theta_{\mathcal{R}}$) or equation (3) (for $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$), one can show that $P_{\text{err}}(\phi_{\text{ML}})$ tends to zero as n tends to infinity. We sketch the analysis here, and defer the detailed proof to the full version [1].

For any $\xi \in \Xi$, let $L(\xi) \triangleq -\log \mathbb{P}_{\xi}(V^{\Omega}, G_1, G_2)$ be the negative log-likelihood of ξ. The estimation rule of ϕ_{ML} is

$$\widehat{\xi} = \phi_{\text{ML}}(V^{\Omega}, G_1, G_2) = \underset{\xi \in \Xi}{\operatorname{argmin}}\ L(\xi).$$

Even though it may not be *a priori* clear, it turns out that the probabilities of error for different ground truths $\xi^* \in \Xi$ with different sizes of $\mathcal{A}_0$ and $\mathcal{R}_0$ are exactly the same. Hence,

$$\begin{aligned} P_{\text{err}}(\phi_{\text{ML}}) &= \mathbb{P}_{\xi^*}(\phi_{\text{ML}}(V^{\Omega}, G_1, G_2) \neq \xi^*) \\ &\le \sum_{\xi \in \Xi : \xi \neq \xi^*} \mathbb{P}_{\xi^*}(L(\xi) \le L(\xi^*)), \end{aligned} \tag{4}$$

where ξ^* can be chosen arbitrarily and (4) follows from the *union bound*. Consider a specific $\xi \neq \xi^*$, and we aim to calculate the probability that $L(\xi) \le L(\xi^*)$. We define $k_1 \triangleq |\xi_{\mathcal{M}} \setminus \xi^*_{\mathcal{M}}| = |\xi_{\mathcal{W}} \setminus \xi^*_{\mathcal{W}}|$ as the amount of overlap between the communities of users in ξ^* and ξ, and define $k_2 \triangleq |\xi_{\mathcal{A}} \setminus \xi^*_{\mathcal{A}}| = |\xi_{\mathcal{R}} \setminus \xi^*_{\mathcal{R}}|$ in a similar way. Let $t_{\mathcal{A}}$ be the number of movies that belong to both $\xi^*_{\mathcal{A}}$ and $\xi_{\mathcal{A}}$, but only one out of $\xi^*_{\mathcal{A}}$ and $\xi_{\mathcal{A}}$ is atypical, i.e., $t_{\mathcal{A}} \triangleq \left|\{j \in [m] : j \in (\xi^*_{\bar{\mathcal{A}}} \cap \xi_{\mathcal{A}_0}) \cup (\xi^*_{\mathcal{A}_0} \cap \xi_{\bar{\mathcal{A}}})\}\right|$. Similarly, let $t_{\mathcal{R}} \triangleq |\{j \in [m] : j \in (\xi^*_{\bar{\mathcal{R}}} \cap \xi_{\mathcal{R}_0}) \cup (\xi^*_{\mathcal{R}_0} \cap \xi_{\bar{\mathcal{R}}})\}|$. With some calculations deferred to [1] and by applying the *Chernoff bound* $\mathbb{P}(X > \kappa) \le \min_{t>0} e^{-t\kappa} \cdot \mathbb{E}(e^{tX})$ with $t = \frac{1}{2}$, we note that $\mathbb{P}_{\xi^*}(L(\xi) \le L(\xi^*))$ depends only on the four parameters $k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}$ defined above. Hence, we partition $\xi \in \Xi \setminus \{\xi^*\}$ into different *type classes* parameterized by $k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}$, which are further denoted by $\Xi_{\xi^*}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})$. Note that all the elements ξ in one type class have the same probability of error, denoted by $P_{\text{err}}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})$. Let $\mathcal{T}$ be the set of valid $(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})$-tuples satisfying $k_1 \in [0 : \frac{n}{4}], k_2 \in [0 : \frac{m}{4}], t_{\mathcal{A}\mathcal{A}} \in [0 : \frac{m}{2} - k_2], t_{\mathcal{R}\mathcal{R}} \in [0 : \frac{m}{2} - k_2]$, and $(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}) \neq (0, 0, 0, 0)$. Then, (4) can be expressed as

$$\sum_{(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}) \in \mathcal{T}} |\Xi_{\xi^*}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})| \cdot P_{\text{err}}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}).$$

For each $(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}}) \in \mathcal{T}$, we calculate the number of ξ in the type class $\Xi_{\xi^*}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})$ and its corresponding $P_{\text{err}}(k_1, k_2, t_{\mathcal{A}}, t_{\mathcal{R}})$. As expected, the most delicate analysis occurs when both k_1 and k_2 are small, since instances ξ that are closer to the ground truth ξ^* are likelier to cause errors. Nonetheless, it turns out that as long as p exceeds the lower bound in equation (2) or equation (3), one can ensure that $\lim_{n \to \infty} P_{\text{err}}(\phi_{\text{ML}}) = 0$. This completes the proof sketch.

V. Conclusion and Future Directions

This paper investigates a novel community recovery problem based on a partially observed rating matrix and two-sided graph side-information. We quantify the gains due to graph side-information; in particular, there exists a certain regime in which simultaneously observing two pieces of graph side-information is critical to reduce the optimal sample probability. In addition to the achievability result, we also develop a converse in [1] showing that for any $\epsilon > 0$, the probability of error of any estimator must tend to one if (i) $\theta_{\mathcal{A}} \neq \theta_{\mathcal{R}}$ and

$$p < \max\left\{\frac{(2(1-\epsilon) - I_1)\log n}{(\nu_{\mathcal{A}} + \nu_{\mathcal{R}})m}, \frac{(1-\epsilon)\log m}{\min\{\nu_{\mathcal{A}}, \nu_{\mathcal{R}}\} \cdot n}, \frac{((1-\epsilon) - I_2)\log m}{2\tau_{\mathcal{A}\mathcal{R}} n}\right\},$$

or (ii) $\theta_{\mathcal{A}} = \theta_{\mathcal{R}}$ and

$$p < \max\left\{\frac{(2(1-\epsilon) - I_1)\log n}{(\nu_{\mathcal{A}} + \nu_{\mathcal{R}})m}, \frac{(1-\epsilon)\log m}{\min\{\nu_{\mathcal{A}}, \nu_{\mathcal{R}}\} \cdot n}\right\}.$$

Note that the achievability and converse results match for a wide range of parameters of interest, and match up to a constant factor of two for the remaining parameter regime.

Finally, we propose two promising directions for future work: (i) In addition to the fundamental limits, another direction that is worth exploring is the algorithmic developments and analyses for such a problem. (ii) It would also be interesting to investigate a more general setting. For instance, users' ratings to movies may not necessarily be binary, and both users and movies may form multiple unequal-sized communities.

REFERENCES

[1] Q. Zhang, V. Y. F. Tan, and C. Suh, "Community detection and matrix completion with two-sided graph side-information," *arXiv preprint arXiv:1912.04099*, 2019.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–71, 1992.

[3] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 287–296.

[4] K. Ahn, K. Lee, H. Cha, and C. Suh, "Binary rating estimation with graph side information," in *Advances in Neural Information Processing Systems*, 2018, pp. 4272–4283.

[5] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 135–142.

[6] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Matrix completion on graphs," *arXiv preprint arXiv:1408.1717*, 2014.

[7] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A graph-based recommender system for digital library," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 2002, pp. 65–73.

[8] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[9] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[10] E. Mossel, J. Neeman, and A. Sly, "Reconstruction and estimation in the planted partition model," *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 431–461, 2015.

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

[12] V. Jog and P.-L. Loh, "Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence," *arXiv preprint arXiv:1509.06418*, 2015.

[13] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4729–4745, 2017.

[14] E. Abbe, "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[15] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.

[16] ——, "Recovering a single community with side information," *arXiv preprint arXiv:1809.01738*, 2018.

[17] ——, "Exact recovery in community detection with continuous-valued side information," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 332–336, 2018.

[18] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *IEEE 13th International Conference on Data Mining*, 2013, pp. 1151–1156.

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

[20] C. Aicher, A. Z. Jacobs, and A. Clauset, "Adapting the stochastic block model to edge-weighted networks," *arXiv preprint arXiv:1305.5782*, 2013.

[21] A. Mazumdar and B. Saha, "Query complexity of clustering with side information," in *Advances in Neural Information Processing Systems*, 2017, pp. 4682–4693.

[22] J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying, "Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, 2014, pp. 29–41.

[23] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Advances in Neural Information Processing Systems*, 2005, pp. 1329–1336.

[24] J. Yoon, K. Lee, and C. Suh, "On the joint recovery of community structure and community features," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 686–694.

[25] C. Jo and K. Lee, "Discrete-valued preference estimation with graph side information," *arXiv preprint arXiv:2003.07040*, 2020.

[26] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Advances in Neural Information Processing Systems*, 2015, pp. 2107–2115.