Binary Rating Estimation with Graph Side Information

Kwangjun Ahn[*]
142nd Military Police Company
Korean Augmentation To the United States Army
kjahnkorea@kaist.ac.kr

Kangwook Lee
School of Electrical Engineering
KAIST
kw1jjang@kaist.ac.kr

Hyunseung Cha
Kakao Brain
tony.cha@kakaobrain.com

Changho Suh
School of Electrical Engineering
KAIST
chsuh@kaist.ac.kr

Abstract

Rich experimental evidences show that one can better estimate users' unknown ratings with the aid of graph side information such as social graphs. However, the gain is not theoretically quantified. In this work, we study the binary rating estimation problem to understand the fundamental value of graph side information. Considering a simple correlation model between a rating matrix and a graph, we characterize the sharp threshold on the number of observed entries required to recover the rating matrix (called the optimal sample complexity) as a function of the quality of graph side information (to be detailed). To the best of our knowledge, we are the first to reveal how much the graph side information reduces sample complexity. Further, we propose a computationally efficient algorithm that achieves the limit. Our experimental results demonstrate that the algorithm performs well even with real-world graphs.

1 Introduction

Recommender systems provide users with appropriate items based on their revealed preference such as ratings and like/dislikes. Due to their wide applicability, recommender systems have received significant attention in machine learning and data mining societies [46, 41, 50, 48, 5, 8, 22, 31]. In addition to the revealed preferences, modern recommender systems also make use of *graph side information* to further improve the performance. For instance, Ma et al. [37] view social networks as user-to-user similarity graph and propose an algorithm that makes use of both revealed ratings and graph side information. As a result, they show that the algorithm can achieve superior performances over those that do not employ social networks. Jamali and Ester [26] demonstrate that an algorithm with graph information can make recommendations for cold start users, whose lack of available rating information precludes the traditional approaches. Similarly, Kalofolias et al. [29] construct an item-to-item similarity graph whose edge weights are computed from the item features, and demonstrate the benefits of exploiting such information.

Apart from the aforementioned, there have been a lot more works that incorporate social graph information in recommender systems; however, few works have been devoted to the theoretical understanding of this problem. In particular, it is widely open as to by how much one can improve the performance with the aid of graph side information. This precisely sets the goal of this paper: we

[*]This work was done when Kwangjun Ahn was with KAIST as a student.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

aim to quantify the gain due to social network information. Specifically, we intend to characterize the optimal sample complexity needed for rating matrix recovery in the presence of the graph side information.

As an initial effort, we focus on a simple setting in which the entries to be estimated are binary. We consider a scenario where one-sided graph information is available, i.e., either user-to-user or item-to-item graph is given. Without loss of generality, we will assume that a user-to-user similarity graph (or so called a social graph) is available.

Consider n users and m items, where each user rates each item either $+1$ (like) or -1 (dislike). The n users are divided into two clusters, and it is assumed that users from the same cluster share their ratings over items. Under this setting, the two types of measurements are available. The first is a partial observation of noisy ratings, and the other is a social graph among the n users, generated as per a celebrated model for random clustered graphs called the stochastic block model (SBM) [21]. Given these, the task is to estimate the ground truth ratings (See Sec. 2 for details). Denote by *the optimal sample complexity* the minimum number of observed ratings required for reliable recovery.

Main contributions. The main contributions of this paper are two-fold. First, under the model of interest, we characterize the optimal sample complexity as a function of the quality of graph information; see Sec. 3 for the quantification of the quality. In particular, we quantify by how much the social network information can reduce sample complexity. Our result demonstrates that the social graph can be as significant as resulting in an order-wise reduction in sample complexity.

The second contribution of this work is to develop an efficient algorithm. The algorithm operates in three stages, and the two types of measurements are used separately during the first two stages and then together in the last stage. We provide a theoretical performance guarantee of this algorithm under the model of interest: we prove that the algorithm reliably recovers the ratings as soon as sample complexity exceeds the optimal sample complexity. We also test the empirical performance of the algorithm to show that it achieves better performances than the state-of-the-art approaches [37, 19] even with real-world graphs, including political blog network [4] and Facebook networks [51].

Related works. Graph side information has been widely used to improve the performance of recommender systems. To begin with, it has been used in matrix factorization-based (MF) approach, which trains the rating matrix from data by assuming that the rating matrix is of low rank. Most works modify the training procedure by adding some regularization terms inspired by the graph side information [26, 37, 34, 10, 29]. Other than regularization techniques, several works modify the existing rating matrix models using graph information [36, 35, 60, 19]. An online version of this problem is also studied [16]. Another popular approach is the one called neighborhood-based approach, in which user's rating information is predicted based on his/her neighborhoods. Several works [38, 18, 54, 24, 25, 55] improve the performance by properly defining neighborhoods using social graphs. Lastly, some recent works propose deep-learning based approaches in which graph information is incorpated into a framework called graph convolutional network [42, 52].

Recently, few works come up with theoretical guarantees for usuefulness of graph side information. Rao et al. [45] provide a consistency guarantee for regularization techniques, demonstrating a gain due to the side information. On the flip side, Chiang et al. [14] consider a model called dirty inductive matrix completion, which incorporates noisy feature-based side information on top of usual low rank matrix completion model. The theoretical guarantee therein show the efficacy of side information unless it is too noisy. However, whether the gains characterized in both works are the maximally achievable remains open. In this work, focusing on a special case, we are able to characterize optimal sample complexity, the maximum gain due to side information.

Moreover, there have been several recent studies that explore the value of side information in the context of *clustering*. In [39], it is proved that similarity information between data points reduces the adaptive query complexity in exact clustering. A feature-based side information was also considered on top of stochastic block model [47, 59]. A different setting is considered in [6] which demonstrates that the k-means problem, which is NP-hard in general, can be solved efficiently with few pairwise queries. In fact, by switching the goal to exactly recover the clusters instead of the ratings, our model can be also seen as a clustering problem with side information. As a consequence of our main result (Theorem 1), we also address this setting; see Corollary 1.

Notation. Let $[n] := \{1, 2, \ldots, n\}$; let $\mathbf{1}_{k,\ell}$ be the $k \times \ell$ all-one matrix; for a graph and two subsets X and Y, let $e(X, Y)$ be the number of edges between X and Y.

2 Model

Consider n users and m items, where m can scale with n. Each user rates each item either $+1$ (like) or -1 (dislike). In real life, people from the same group tend to share their preferences, recommending commodities with each other.[2] In an effort to capture this, we consider a simple setting in which there are two clusters of the same size, and the users from the same cluster share the ratings over items. More precisely, we consider a binary rating matrix $R \in \{+1, -1\}^{n \times m}$ such that half of its rows are u_R and the other half are v_R, for some u_R and v_R, which we call *rating vectors* of R. We denote these two exclusive sets (or clusters) of row indices by A_R and B_R, respectively.

Observation model. Given R, we consider two types of measurements.

1. Partial observation of noisy ratings (N^Ω) : We observe each entry of R independently with probability p, where $0 \leq p \leq 1$. We further assume that our observed entries are noisy: the value of an observed entry can be flipped with probability $\theta \in (0, \frac{1}{2})$. We denote the subset of observed entries by $\Omega \subset [n] \times [m]$. We represent this with an observation matrix N^Ω of size $n \times m$, whose (i, j)-th entry is the noisy observation if $(i, j) \in \Omega$ and 0 otherwise. In other words, $N^\Omega_{ij} \overset{\text{i.i.d.}}{\sim} R_{ij} \cdot \mathsf{Bern}(p) \cdot (1 - 2\mathsf{Bern}(\theta))$.

2. Social graph information (G): The social graph on n users is generated as per the stochastic block model (SBM) [21]. That is, given the two clusters of users A_R and B_R, an edge between each pair of two users i, j is placed with probability α, independently of the others, if they are from the same cluster. If not, the probability of having an edge between them is β, where $\alpha \geq \beta$. Let $G = ([n], E)$ denote the social graph.

Performance metric. Given N^Ω and G, the task of interest is to recover R. The performance of an estimator is measured by the probability that the output of estimator does not coincide with R, namely *the probability of error*. Concretely, we assume that the worst-case matrix is chosen from a collection of rating matrices R' with $\|u_{R'} - v_{R'}\|_0 = \lceil \gamma m \rceil$, where $\gamma \in (0, 1)$ is a fixed constant.

Definition 1. *For a fixed constant $\gamma \in (0, 1)$ and an estimator ψ that outputs a binary rating matrix based on N^Ω and G, the worst-case probability of error $P_e^{(\gamma)}(\psi)$ is defined as* $P_e^{(\gamma)}(\psi) := \max_{R' : \|u_{R'} - v_{R'}\|_0 = \lceil \gamma m \rceil} \Pr\left(\psi(N^\Omega, G) \neq R'\right)$.

Goal. We aim to characterize $p^\star$ such that (i) when p exceeds $p^\star$, $P_e^{(\gamma)}(\psi) \to 0$ as $n \to \infty$ for some estimator ψ; (ii) when p is less than $p^\star$, $P_e^{(\gamma)}(\psi) \not\to 0$ for any ψ. In particular, we aim to characterize $nmp^\star$ that we call *the optimal sample complexity*. Given the fact that nmp is the expected number of observed entries, the optimal sample complexity can be seen as the minimum number of observed entries required for rating recovery in the limit of n.

3 Optimal sample complexity

We characterize the optimal sample complexity as a function of $n, m, \theta, \gamma, \alpha$, and β.

Theorem 1. *Let $\gamma \in (0, 1)$ be fixed. Assume that $m = \omega(\log n)$ and $\log m = o(n)$.*[3] *Then, the following holds for any constant $\epsilon > 0$: if*

$$p \geq \frac{1}{(\sqrt{1-\theta} - \sqrt{\theta})^2} \max\left\{ \frac{(1+\epsilon)\log n - \frac{n}{2}(\sqrt{\alpha} - \sqrt{\beta})^2}{\gamma m}, \frac{(1+\epsilon)2\log m}{n} \right\},$$

then $P_e^{(\gamma)}(\psi) \to 0$ as $n \to \infty$ for some ψ that outputs a binary rating matrix based on N^Ω and G ; conversely, if

$$p \leq \frac{1}{(\sqrt{1-\theta} - \sqrt{\theta})^2} \max\left\{ \frac{(1-\epsilon)\log n - \frac{n}{2}(\sqrt{\alpha} - \sqrt{\beta})^2}{\gamma m}, \frac{(1-\epsilon)2\log m}{n} \right\},$$

then $P_e^{(\gamma)}(\psi) \not\to 0$ as $n \to \infty$ for any ψ.

[2] This tendency, called *homophily*, has been extensively studied in sociology and psychology [40].

[3] We employ these conditions to obtain the large deviation results in the proof (See the supplemental material). Intuitively, these conditions rule out tall and fat matrices, respectively.

Proof: See Sec. 5 for the proof sketch; and see the supplemental material for the full proof . □

Let us interpret Theorem 1. See Table 1 for a summary. In essence, Theorem 1 asserts that the rating recovery is possible if and only if

$$p > p^\star := \frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \max\left\{(\gamma m)^{-1}\log n - (2\gamma m)^{-1} n(\sqrt{\alpha}-\sqrt{\beta})^2,\ n^{-1}2\log m\right\}.$$

For illustrative purpose, we introduce a notation: $I_s := (\sqrt{\alpha}-\sqrt{\beta})^2$. One can interpret I_s as the quality of social graph information. This is because the two-cluster structure becomes more transparent as the gap between α and β gets larger. For instance, when $\alpha = \beta$, i.e., $I_s = 0$, there is no way to distinguish the two clusters. On the other hand, when $\alpha = 1$ and $\beta = 0$ (or $\alpha = 0$ and $\beta = 1$), i.e., $I_s = 1$, the cluster structure is straightforward from G. Note that the notation I_s is also employed in the context of community recovery under SBM, in which the fundamental limit for exact recovery is shown to be $I_s > 2\frac{\log n}{n}$ [1].

First, consider $I_s = 0$. In this case, the optimal sample complexity $nmp^\star$ is

$$\frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \max\{\gamma^{-1} n\log n, 2m\log m\}. \tag{1}$$

Remark 1. *Note that the fundamental limit decreases in γ. This tendency is intuitive. To see this, focus on the noiseless setting ($\theta = 0$). Let us classify entries of the rating matrix into two types: (i) the entries on the columns where the ratings of two groups coincide; and (ii) the other entries. Unlike the first type of entries, the second type of entries have possibility to reveal significant information on clusters: when we observe two users' different ratings on the same item, it can be immediately concluded that the two users belong to different clusters. Hence, the second type of entries are more informative. As there are γ-fraction of entries of the second type, the chance of getting more informative entries increases in γ. Thus, the required sample complexity decreases in γ.*

We now turn to the case $I_s \neq 0$. In this case, the optimal sample complexity $nmp^\star$ is

$$\frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \max\{\gamma^{-1} n\log n - (2\gamma)^{-1} n^2 I_s,\ 2m\log m\}. \tag{2}$$

On the one hand, this result suggests that the social graph information does not help rating recovery when the number of users is relatively smaller than that of items. More precisely, when $2n\log n \leq 4\gamma m\log m$, both (1) and (2) are equal to $2m\log m/(\sqrt{1-\theta}-\sqrt{\theta})^2$.

On the other hand, when $2n\log n > 4\gamma m\log m$, i.e., (1) is equal to $\gamma^{-1} n\log n/(\sqrt{1-\theta}-\sqrt{\theta})^2$, the result implies that the social graph information does help rating recovery. Below, we examine the amount of reduction in sample complexity as a function of I_s. For simplicity, we focus on a setting in which $\theta = 0$ and $\gamma = \frac{1}{2}$, i.e., (1) is equal to $2n\log n$.

We first consider the case of $n^2 I_s < 2n\log n - 2m\log m$, i.e., (2) being equal to $2n\log n - n^2 I_s$. In this case, sample complexity is reduced by $n^2 I_s$. That being said, there is no asymptotical gain unless $n^2 I_s = \Omega(n\log n)$. On the other hand, when $n^2 I_s = \Omega(n\log n)$, this reduction can be as significant as resulting in an order-wise reduction in sample complexity. To see this clearly, consider two scenarios: $n = 2m$ and $n = m^2$. Also see Fig. 1.

Example 1 ($n = 2m$)**.** *Note that $n^2 I_s < 2n\log n - 2m\log m$ if and only if $I_s < \log(2n)/n$. When $I_s = c\log(2n)/n$ for $0 < c < 1$, the optimal sample complexity is $2n\log n - cn\log 2n = (2-c)n\log n - cn\log 2$, which is (asymptotically) lower than* (1) *by a multiplicative factor of $\frac{c}{2}$.*

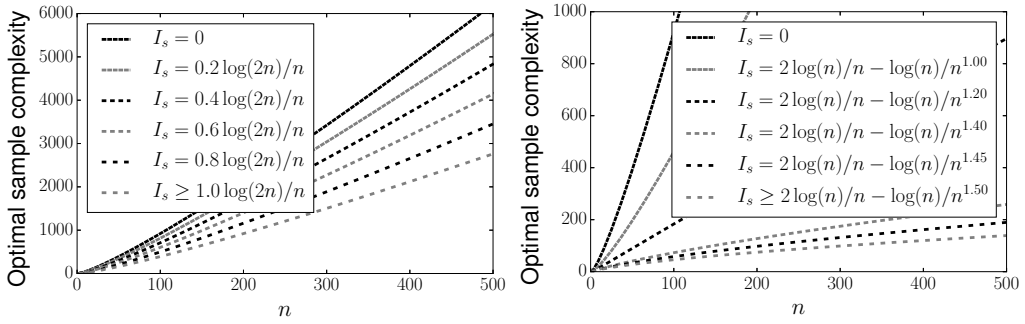
Example 2 ($n = m^2$)**.** *Note that $n^2 I_s < 2n\log n - 2m\log m$ if and only if $I_s < 2\log n/n - \log n/n^{1.5}$. When $I_s = 2\log n/n - \log n/n^c$ for $1 < c < 1.5$, the optimal sample complexity is $n^{2-c}\log n$, which shows an order-wise reduction in sample complexity.*

Remark 2. *Example 2 justifies the observation made by Jamali and Ester [26] that graph side information can help predict ratings for cold start users. More specifically, note that when $c = 1.4$, most users are cold start users as the average number of observed ratings per user is $n^{-0.4}\log n$.*

For the case of $n^2 I_s \geq 2n\log n - 2m\log m$, (2) $= 2m\log m$ no matter how large I_s is. This implies that the gain due to side information is saturated.

Table 1: **Summary of the gain in sample complexity.** Depending on the quality of graph information $I_s := (\sqrt{\alpha} - \sqrt{\beta})^2$, the gain in sample complexity can be summarized as follows. First, graph information does not help unless the number of users is relatively larger than that of items ($2n \log n \geq 4\gamma m \log m$). When it helps, the efficacy of the information depends on its quality: When $n^2 I_s = o(n \log n)$, social network is too noisy, and hence does not help rating recovery. When $I_s = \Omega(\log n/n)$, the minimum sample complexity is a decreasing function in I_s. In other words, as the quality of social network increases, the minimum sample complexity decreases. However, the gain of side-information is characterized as diminishing returns: When I_s is larger than a certain threshold, the minimum sample complexity stops decreasing. Note that this does not mean that graph information does not help: It helps but its gain is saturated.

Value of $n^2 I_s$		
$o(n \log n)$	$< 2n \log n - 4\gamma m \log m$	$\geq 2n \log n - 4\gamma m \log m$
no asymptotical gain	gain increases in I_s	gain is saturated.



(a) $n = 2m$ (b) $n = m^2$

Figure 1: **Significant reduction in sample complexity due to social graph.** We illustrate Theorem 1 for two cases: $n = 2m$ and $n = m^2$. For $n = 2m$, the optimal sample complexity is reduced by some multiplicative factors, and for $n = m^2$, there is an order-wise reduction in sample complexity. See Example 1 and 2 for details.

Implications on community recovery. Switching the goal of our model to exactly recover A_R and B_R instead of R, our model can cover a community recovery problem with some side information with respect to ratings. The proof of Theorem 1 suggests the fundamental limit on I_s for exact recovery:

Corollary 1. *Suppose we wish to exactly recover the clusters A_R and B_R (up to a flip) instead of R (we should also modify the definition of $P_e^{(\gamma)}(\psi)$ accordingly). Then, the following holds for any constant $\epsilon > 0$: $P_e^{(\gamma)}(\psi) \to 0$ as $n \to \infty$ for some ψ that outputs two equal-sized clusters based on N^Ω and G whenever $I_s + \frac{2\gamma mp(\sqrt{1-\theta}-\sqrt{\theta})^2}{n} \geq (1+\epsilon)\frac{2\log n}{n}$; conversely, $P_e^{(\gamma)}(\psi) \not\to 0$ as $n \to \infty$ for any ψ if $I_s + \frac{2\gamma mp(\sqrt{1-\theta}-\sqrt{\theta})^2}{n} \leq (1-\epsilon)\frac{2\log n}{n}$.*

This result implies that when $p \neq 0$, the amount of graph information I_s required for cluster recovery reduces from $\frac{2\log n}{n}$ [1] to $\max\left\{\frac{2\log n}{n} - \frac{2\gamma mp(\sqrt{1-\theta}-\sqrt{\theta})^2}{n}, 0\right\}$.

4 Proposed algorithm

In this section, we propose an efficient rating estimation algorithm, which achieves the fundamental limit characterized in Theorem 1. We note that our algorithm can also be applied to a general setting (not limited to the simple two-cluster setting described earlier), although the theoretical guarantee for the setting is not provided. This will be clearer while describing our algorithm.

Algorithm description. The algorithm works in three stages. The inputs of the algorithm are N^Ω, G, the number of clusters k and hyperparameters $c_1, c_2 > 0$.[4]

[4] One can always use validation data to tune c_1 and c_2. For the case of two equal-sized communities, we characterized the optimal choice of c_1 and c_2 in Theorem 2.

Stage 1 (Partial recovery of clusters): First, we run a spectral method [17] on G. Let $A_1^{(0)}, A_2^{(0)}, \ldots, A_k^{(0)}$ be the output of the clustering. Note that other clustering algorithms such as other spectral methods [2, 15, 33], nonbacktracking matrix based methods [32], semidefinite programming (SDP) [27], and belief propagation (BP) variants [43] can be employed for this stage.

Stage 2 (Recovery of rating vectors): Next, for each j, we recover the rating vector of the cluster $A_j^{(0)}$ using the observed ratings. For each item i, we collect the observed ratings of the item by the users in $A_j^{(0)}$. Among the collected ratings, we find the one that appears most frequently; let the rating be $u_i^{(j)}$. The rating vector is then defined as $u^{(j)} = \left[u_i^{(j)}\right]_{i=1}^{m}$.

Stage 3 (Local refinement of clusters): The last stage iteratively refines the clusters $A_j^{(0)}$'s using G, N^{Ω} and $u^{(j)}$'s. This stage consists of T times of refinement steps, and in each step, the clusters are updated as follows. Let $A_1^{(t-1)}, A_2^{(t-1)}, \ldots, A_k^{(t-1)}$ be the outcome of $(t-1)$-th refinement step for some $t = 1, 2, \ldots, T$. Then, for each $i = 1, 2, \ldots, n$, we put user i to $A_{j^*}^{(t)}$, which is the one among $A_j^{(t-1)}$'s that gives the maximum value of

$$c_1 \cdot e\left(\{i\}, A_j^{(t-1)}\right) - c_2 \cdot \bar{e}\left(\{i\}, A_j^{(t-1)}\right) + \Pi_i(u^{(j)}).$$

Here, $\bar{e}(\{i\}, A_j^{(t-1)}) := |A_j^{(t-1)}| - e(\{i\}, A_j^{(t-1)})$ indicates the number of non-edges between i and $A_j^{(t-1)}$; $\Pi_i(u^{(j)})$ is the number of user i's observed ratings which coincide with that of $u^{(j)}$.

Lastly, the algorithm outputs $\widehat{R}$ where i-th row is $u^{(j)}$ whenever $i \in A_j^{(T)}$.

Remark 3 (Update rule in Stage 3)**.** *For a cluster $A_j^{(t-1)}$ and its rating vector $u^{(j)}$, the term $c_1 e\left(\{i\}, A_j^{(t-1)}\right) + \Pi_i(u^{(j)})$, which is the sum of the first and the third terms of the update rule, can be seen as a measure of* fitness *of user i with respect to the cluster $A_j^{(t-1)}$. This is because i is more likely to belong to $A_j^{(t-1)}$ when there are more edges between i and $A_j^{(t-1)}$ and more observed ratings consistent with $u^{(j)}$. The number of non-edges is subtracted from the term to minimize the effect of cluster size as a large cluster tends to have more edges to users.*

Remark 4 (Non-agnostic refinement rule)**.** *When the algorithm knows the size of each cluster, we modify the definition of $\bar{e}$ by replacing the terms $|A_j^{(t-1)}|$'s with the actual sizes $|A_j|$'s. In particular, when the algorithm knows that the clusters are of equal size, we obtain the following refinement rule: we put user i to $A_{j^*}^{(t)}$, where $j^* = \arg\max_{j \in [k]} \left[c_1' \cdot e\left(\{i\}, A_j^{(t-1)}\right) + \Pi_i(u^{(j)})\right]$ for some hyperparameter $c_1' > 0$.*

Remark 5. *Note that this algorithm can be applied to a general setting where (i) rating type is not limited to binary; (ii) the number of clusters can be larger than 2; and (iii) the clusters are of unequal sizes.*

Remark 6. *The proposed algorithm is inspired by a general paradigm in solving non-convex problems: first obtain a decent initial estimate and iteratively refine the estimate to reach the global optimum. This paradigm has been employed in various contexts, including matrix completion [30, 23], phase retrieval [44, 11], robust PCA [56], community recovery [2, 17, 57, 12], EM-algorithm [7], and rank aggregation [13]. Moreover, we note that spectral algorithms have been also used in rating estimation in the context of crowdsourcing [49, 53].*

One important aspect of this algorithm is its low computational complexity. Spectral methods can be run within $O(|E| \log n)$ time using the power method [9]. Stage 2 requires a single pass of all observed ratings, which amounts to $O(|\Omega|)$ time. As for Stage 3, a single individual update of user i entails reading of i-th row and edges connected to user i. Assuming a proper data structure, each iteration requires $O(|E| + |\Omega|)$. Hence, Stage 3 can be done within $O(|E|T + |\Omega|T)$. Overall, the proposed algorithm can be performed within $O(|E|T + |E| \log n + |\Omega|T)$.

Theoretical performance guarantee. To investigate theoretical guarantees of the proposed algorithm, we focus on the model in Sec. 2.

Theorem 2. *Let R be any binary rating matrix with $\|u_R - v_R\|_0 = \gamma m$ for some $\gamma \in (0,1)$. In addition to the assumptions in Theorem 1, assume further that $m = O(n)$ and $I_s = \omega(1/n)$.*[5] *Suppose that there exists $\epsilon > 0$ such that the sufficient condition in Theorem 1 holds. Then with probability approaching 1 as $n \to \infty$, the proposed algorithm with the knowledge that the two clusters are of equal size (i.e., using the refinement rule in Remark 4) exactly recovers R under the settings $T = O(\log n)$ and $c_1' = \log\left(\frac{\hat{\alpha}(1-\hat{\beta})}{\hat{\beta}(1-\hat{\alpha})}\right) / \log\left(\frac{1-\hat{\theta}}{\hat{\theta}}\right)$. Here, $\hat{\alpha} := \frac{e(A_1(0),A_1(0))+e(A_2(0),A_2(0))}{2\binom{n/2}{2}}$ and $\hat{\beta} := 4\frac{e(A_1^{(0)},A_2^{(0)})}{n^2}$ are estimations of model parameters α and β after Stage 1, and $\hat{\theta}$ is an estimation of θ after Stage 2 defined by the fraction of observed ratings which are different from the corresponding entries of the rating matrix defined by clusters $A_1^{(0)}, A_2^{(0)}$ and rating vectors $u^{(1)}, u^{(2)}$.*

Proof: Due to space limitation, we defer the proof to the supplemental material. □

Theorem 2 implies that the proposed algorithm with proper hyperparameter choices achieves the fundamental limit in Theorem 1 except for the case of scarce social graph information ($I_s = O(1/n)$).

5 Proof outline of Theorem 1

We sketch the proof while deferring the full proof to the supplemental material. Let $I_r := p(\sqrt{1-\theta} - \sqrt{\theta})^2$. Using the notations of I_r and I_s, one can succinctly represent the sufficient condition and the necessary condition claimed in Theorem 1. For instance, the sufficient condition reads

$$\frac{1}{2}nI_s + \gamma m I_r \geq (1+\epsilon)\log n \qquad \text{and} \qquad \frac{1}{2}nI_r \geq (1+\epsilon)\log m\,.$$

We next introduce a few more notations that will be used throughout the proof. Let $\mathcal{C}^{(\gamma)}$ be the collection of rating matrices R such that $\|u_R - v_R\|_0 = \gamma m$ (Here and below, we treat γm as an integer for notational simplicity); let $R^{(\gamma)} := \left[\begin{array}{c|c} +\mathbf{1}_{n/2,(1-\gamma)m} & +\mathbf{1}_{n/2,\gamma m} \\ \hline +\mathbf{1}_{n/2,(1-\gamma)m} & -\mathbf{1}_{n/2,\gamma m} \end{array}\right] \in \mathcal{C}^{(\gamma)}$ (i.e., $A_{R^{(\gamma)}} = [\frac{n}{2}]$, $B_{R^{(\gamma)}} = [n] \setminus [\frac{n}{2}]$, $u_{R^{(\gamma)}} = +\mathbf{1}_{1,m}$ and $v_{R^{(\gamma)}} = [+\mathbf{1}_{1,(1-\gamma)m} \mid -\mathbf{1}_{1,\gamma m}]$). Lastly, let ψ_{ML} be the maximum likelihood estimator (output being not constrained to $\mathcal{C}^{(\gamma)}$) and $\mathsf{L}(\cdot)$ be the likelihood function.

Achievability: It is enough to show that $P_e^{(\gamma)}(\psi_{\text{ML}}) \to 0$. By symmetry, we fix the ground truth rating matrix to be $R^{(\gamma)}$. Note that the event "$\psi_{\text{ML}}(N^\Omega, G) \neq R^{(\gamma)}$" happens only if $\mathsf{L}(X) \leq \mathsf{L}(R^{(\gamma)})$ for some binary rating matrix X. Hence, by the union bound,

$$P_e^{(\gamma)}(\psi_{\text{ML}}) \leq \sum_{X \neq M} \Pr\left(\mathsf{L}(X) \leq \mathsf{L}(R^{(\gamma)})\right)\,. \tag{3}$$

To enumerate all rating matrices different from $R^{(\gamma)}$, we define $\mathcal{X}(k, a_1, a_2, b_1, b_2)$ to be the class of rating matrices X's such that (i) $|A_X \setminus A_{R^{(\gamma)}}| = |B_X \setminus B_{R^{(\gamma)}}| =: k$; (ii) u_X differs from $u_{R^{(\gamma)}}$ at a_1 many coordinates among the first $(1-\gamma)m$ coordinates and at a_2 many coordinates among the next γm coordinates; and (iii) v_X differs from $v_{R^{(\gamma)}}$ at b_1 many coordinates among the first $(1-\gamma)m$ coordinates and at b_2 many coordinates among the next γm coordinates. Note that if X_1 and X_2 belong to the same class, then

$$\Pr\left(\mathsf{L}(X_1) \leq \mathsf{L}(M)\right) = \Pr\left(\mathsf{L}(X_2) \leq \mathsf{L}(M)\right)$$

as the two events are statistically identical. Let $\mathcal{I}$ be the range of index, i.e., collection of tuples $(k, a_1, a_2, b_1, b_2) \neq (0,0,0,0,0)$ such that $0 \leq k \leq n/4$ and $0 \leq a_1, b_1 \leq (1-\gamma)m$ and $0 \leq a_2, b_2 \leq \gamma m$. Note that $k \leq n/4$ is sufficient as one can switch the role of u_X and v_X.

For each 5-tuple $z \in \mathcal{I}$, let X_z be a binary rating matrix in $\mathcal{X}(z)$. With this enumeration, the right hand side of (3) becomes $\sum_{z\in\mathcal{I}} |X(z)| \Pr\left(\mathsf{L}(X_z) \leq \mathsf{L}(M)\right)$.

Let $z = (k, a_1, a_2, b_1, b_2)$. To upper bound $\Pr\left(\mathsf{L}(X_z) \leq \mathsf{L}(M)\right)$, we developed a large deviation result building upon the techniques in [28, 58]: for $z = (k, a_1, a_2, b_1, b_2)$,

$$\Pr\left(\mathsf{L}(X_z) \leq \mathsf{L}(M)\right) \leq e^{-2(\frac{n}{2}-k)kI_s - \mathcal{D}_z I_r}\,, \tag{4}$$

[5] Note that the condition $m = O(n)$ is for reliable estimation of parameters α, β, θ, and hence can be removed when the parameters are known.

where $\mathcal{D}_z := k \cdot \{a_1 + b_1 + (\gamma m - a_2) + (\gamma m - b_2)\} + \left(\frac{n}{2} - k\right) \cdot (a_1 + a_2 + b_1 + b_2)$; we refer readers to the supplemental material for details. Let $S(z) := |X(z)| e^{-2(\frac{n}{2}-k)kI_s - \mathcal{D}_z I_r}$. Then, the last upper bound is bounded by $\sum_{z \in \mathcal{I}} S(z)$.

In the supplemental material, we show that $S(z)$'s for z with at least one large coordinate are negligible. More precisely, for $\mathcal{L} := \{(k, a_1, a_2, b_1, b_2) \,:\, k < \delta n \text{ and } a_1, a_2, b_1, b_2 < \delta m\}$ (where δ is a sufficiently small quantity), $\sum_{\mathcal{I} \setminus \mathcal{L}} S(z)$ is negligible compared to $\sum_{\mathcal{I} \cap \mathcal{L}} S(z)$. The rationale behind this is that when $z \in \mathcal{I} \setminus \mathcal{L}$, $S(z)$ becomes a very small quantity as either $2(\frac{n}{2} - k)k$ or $\mathcal{D}_z$ becomes very large.

Hence, it suffices to focus on $\sum_{\mathcal{I} \cap \mathcal{L}} S(z)$. As $k \ll n$ and $a_1, a_2, b_1, b_2 \ll m$ when $z \in \mathcal{I} \cap \mathcal{L}$, one can approximate $2(\frac{n}{2} - k)k \approx nk$ and $\mathcal{D}_z \approx 2\gamma km + \frac{n}{2}(a_1 + a_2 + b_1 + b_2)$. By definition,

$$|\mathcal{X}(k, a_1, a_2, b_1, b_2)| = \binom{\frac{n}{2}}{k}^2 \binom{(1-\gamma)m}{a_1} \binom{(1-\gamma)m}{b_1} \binom{\gamma m}{a_2} \binom{\gamma m}{b_2} \leq n^{2k} m^{a_1 + a_2 + b_1 + b_2} .$$

This together with the above approximation yields

$$\begin{aligned} S(z) &\leq e^{2k \log n + (a_1 + a_2 + b_1 + b_2) \log m} e^{-2(\frac{n}{2} - k)kI_s - \mathcal{D}_z I_r} \\ &\approx e^{-k \cdot (nI_r + 2\gamma m I_s - 2 \log n) - (a_1 + a_2 + b_1 + b_2) \cdot (\frac{n}{2} I_r - \log m)} \leq (n^{-2\epsilon})^k (m^{-\epsilon})^{(a_1 + a_2 + b_1 + b_2)} , \end{aligned}$$

where the last inequality is due to $\frac{1}{2} n I_s + \gamma m I_r \geq (1 + \epsilon) \log n$ and $\frac{1}{2} n I_r \geq (1 + \epsilon) \log m$. This justifies $\sum_{z \in \mathcal{I} \cap \mathcal{L}} S(z) \to 0$.

Converse: *Step 1 (ML as an optimal estimator):* Consider the maximum likelihood estimator $\psi_{\text{ML}}|_{\mathcal{C}^{(\gamma)}}$ whose output is constrained in $\mathcal{C}^{(\gamma)}$. It can be proven that

$$\inf_{\psi} P_e^{(\gamma)}(\psi) \geq P_e^{(\gamma)}(\psi_{\text{ML}}|_{\mathcal{C}^{(\gamma)}}) .$$

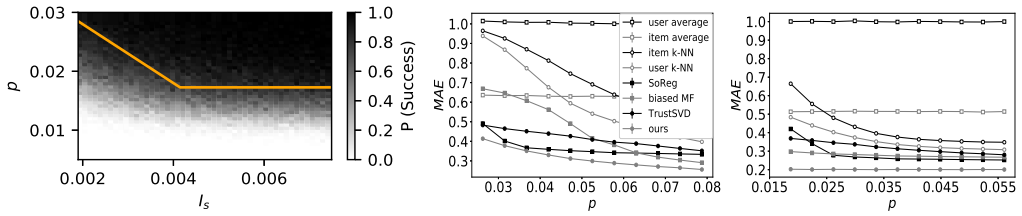
See the supplemental material for details. Hence, it is enough to show $P_e^{(\gamma)}(\psi_{\text{ML}}|_{\mathcal{C}^{(\gamma)}}) \not\to 0$. By symmetry, we fix the ground truth rating matrix to be $R^{(\gamma)}$.

Step 2 (Genie-aided ML estimators): We consider *genie-aided* ML estimators, in which the genie helps the estimator by telling the answer is one of few candidates within $\mathcal{C}^{(\gamma)}$. Owing to the notation $\mathcal{X}(\cdot, \cdot, \cdot, \cdot, \cdot)$ in the achievability proof, two different kinds of genie-aided estimators are examined: $\psi_{\text{ML}}^{(1)}$ is given with the information that the ground truth belongs to $R^{(\gamma)} \cup \mathcal{X}(0, 1, 1, 0, 0)$; $\psi_{\text{ML}}^{(2)}$ is given with the information that the ground truth belongs to $R^{(\gamma)} \cup \mathcal{X}(1, 0, 0, 0, 0)$.

Step 3 (Analysis of genie-aided estimators): We prove (i) $\psi_{\text{ML}}^{(1)}$ fails if $\frac{1}{2} n I_r \leq (1 - \epsilon) \log m$ and (ii) $\psi_{\text{ML}}^{(2)}$ fails if $\frac{1}{2} n I_s + \gamma m I_r \leq (1 - \epsilon) \log n$. Here, we provide the proof sketch of (i): we remark that the proof of (ii) is trickier and it requires some combinatorial properties of random graphs. Note that if the likelihood $\mathsf{L}(X)$ for some $X \in \mathcal{X}(0, 0, 0, 1, 1)$ is less than or equal to $\mathsf{L}(R^{(\gamma)})$, then $\psi_{\text{ML}}^{(1)}$ fails with probability at least $1/2$. Hence, it is enough to show that with probability approaching 1, there exists $X \in \mathcal{X}(0, 0, 0, 1, 1)$ such that $\mathsf{L}(X) \leq \mathsf{L}(R^{(\gamma)})$, or equivalently (by taking complement), $\Pr\left(\bigcap_{X \in \mathcal{X}(0,0,0,1,1)} \left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]\right) \to 0$. On the other hand, some difficulties arise while analysing the last probability as the events $\left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]_{X \in \mathcal{X}(0,0,0,1,1)}$ are not mutually independent. A trick avoiding this issue is to show that the last probability is bounded by $\Pr\left(\bigcap_{X \in \mathcal{X}(0,0,0,1,0)} \left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]\right) + \Pr\left(\bigcap_{X \in \mathcal{X}(0,0,0,0,1)} \left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]\right)$. The analysis is now tractable as the collections of events $\left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]_{X \in \mathcal{X}(0,0,0,1,0)}$ and $\left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]_{X \in \mathcal{X}(0,0,0,0,1)}$ are both mutually independent. Now, we conclude the proof by using the reverse direction of the bound (4) (the bound (4) is indeed tight and the reverse direction also holds up to a constant factor; see the supplemental material): For instance,

$$\Pr\left(\bigcap_{X \in \mathcal{X}(0,0,0,1,0)} \left[\mathsf{L}(X) > \mathsf{L}(R^{(\gamma)})\right]\right) \leq (1 - e^{-\frac{n}{2} I_r})^{|\mathcal{X}(0,0,0,1,0)|} \leq e^{-(1-\gamma) m e^{-\frac{n}{2} I_r}} ,$$

where the last inequality is due to the inequality $1 - x \leq e^{-x}$ and $|\mathcal{X}(0, 0, 0, 1, 0)| = (1 - \gamma)m$; the last term goes to zero when $\frac{1}{2} n I_r \leq (1 - \epsilon) \log m$.



(a) Synthetic data (b) Poliblog (2 clusters) (c) Facebook (6 clusters)

Figure 2: (a) The level of darkness depicts the empirical success rate, and the orange line reflects the optimal sample complexity due to Theorem 1. A sharp transition in darkness near the line corroborates Theorem 1. (b&c) Performance comparison of algorithms on real datasets. Our algorithm shows better performance than the other algorithms on every data set, demonstrating practicality of our approach.

6 Experiments

We first conduct an experiment to corroborate Theorem 1. We consider a setting where $n = 2000$ users and $m = 1000$ items. The synthetic data is generated as per the model in Sec. 2.[6] For each triple (p, α, β), an empirical success rate of the proposed algorithm is measured over 100 random trials. Figure 2a shows the result, where the empirical success rate is depicted by the level of darkness. The orange (solid) line reflects the optimal sample complexity due to Theorem 1. The phase transition occurs near the orange line, corroborating Theorem 1.

We next show that our proposed algorithm performs well even with real-world graphs. On top of the real graphs (political blog network [4] and Facebook networks [51]), we synthesize ratings as per our model. For the performance metric, we use mean absolute error (MAE): $\mathbb{E}[|\widehat{R}_{ij} - R_{ij}|]$, where the expectation is over the observed ratings. We then compare the performance of our algorithm with 7 well-known recommendation algorithms. Specifically, we compare the performance of our algorithm with 7 recommendation algorithms.[7] Reported in Figure 2b, 2c are the performances of rating estimation algorithms on real graph data. Our algorithm shows better performance than the other algorithms, showing the practicality of our approach.

7 Conclusion

Motivated by the lack of study in quantifying the value of social graph information in recommender systems, this work characterized the optimal sample complexity of the rating recovery problem with social graph information. We also proposed an efficient rating estimation algorithm that provably achieves the optimal sample complexity.

This paper comes with some limitations in characterizing sample complexity for more general models. We hope restrictive assumptions considered in this paper, such as binary ratings and rating being shared across the same group, be relaxed in the future endeavors. In particular, it would be interesting to characterize the optimal sample for feature-based side information models [14, 45]. Moreover, as in the case of community detection, sample complexity for partial recovery [3] might be more desirable in practice over our exact recovery setting.

References

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

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *FOCS*, pages 670–688. IEEE, 2015.

[3] Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures in the general stochastic block model. *To Appear in Communications on Pure and Applied Mathematics*, 2017.

[6] We set $\theta = 0.1$ and $\gamma = 0.5$. We vary α and p while fixing $\beta = \frac{\log n}{n}$.

[7] The recommendation algorithms include item average, user average, item k-Nearest Neighbor (NN), user k-NN, biased matrix factorization [31], matrix factorization with social regularization (SoReg) [37], and TrustSVD [19]. We adopt the implementations from LibRec, an open-sourced Java library for recommendation systems [20].

[4] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[5] Deepak Agarwal and Bee-Chung Chen. flda: matrix factorization through latent dirichlet allocation. In *WSDM*, pages 91–100. ACM, 2010.

[6] Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *Advances in neural information processing systems*, pages 3216–3224, 2016.

[7] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 2017.

[8] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *SIGKDD*, pages 95–104. ACM, 2007.

[9] Christos Boutsidis, Prabhanjan Kambadur, and Alex Gittens. Spectral clustering via the power method-provably. In *ICML*, pages 40–48, 2015.

[10] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[11] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[12] Yuxin Chen, Govinda Kamath, Changho Suh, and David Tse. Community recovery in graphs with locality. In *ICML*, 2016.

[13] Yuxin Chen and Changho Suh. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *ICML*, pages 371–380, 2015.

[14] Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems*, pages 3447–3455, 2015.

[15] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *COLT*, pages 391–423, 2015.

[16] Symeon Chouvardas, Mohammed Amin Abdullah, Lucas Claude, and Moez Draief. Robust online matrix completion on graphs. In *ICASSP*, pages 4019–4023. IEEE, 2017.

[17] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *JMLR*, 2017.

[18] Jennifer Golbeck, James Hendler, et al. Filmtrust: Movie recommendations using trust in web-based social networks. In *IEEE CCNC*, pages 282–286, 2006.

[19] G. Guo, J. Zhang, and N. Yorke-Smith. Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings. In *AAAI*, 2015.

[20] Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. Librec: A java library for recommender systems. In *UMAP Workshops*, volume 4, 2015.

[21] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[22] Michael Jahrer, Andreas Töscher, and Robert Legenstein. Combining predictions for accurate recommender systems. In *SIGKDD*, pages 693–702. ACM, 2010.

[23] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674. ACM, 2013.

[24] Mohsen Jamali and Martin Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *SIGKDD*, pages 397–406. ACM, 2009.

[25] Mohsen Jamali and Martin Ester. Using a trust network to improve top-n recommendation. In *RecSys*, pages 181–188. ACM, 2009.

[26] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, pages 135–142. ACM, 2010.

[27] Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *PNAS*, 113(16):E2218–E2223, 2016.

[28] Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

[29] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*, 2014.

[30] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 2010.

[31] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*, 2008.

[32] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *PNAS*, 2013.

[33] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[34] Wu-Jun Li and Dit Yan Yeung. Relation regularized matrix factorization. In *IJCAI*, page 1126, 2009.

[35] Hao Ma, Irwin King, and Michael R Lyu. Learning to recommend with social trust ensemble. In *SIGIR*, pages 203–210. ACM, 2009.

[36] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, pages 931–940. ACM, 2008.

[37] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. Recommender systems with social regularization. In *WSDM*, pages 287–296. ACM, 2011.

[38] Paolo Massa and Paolo Avesani. Controversial users demand local trust metrics: An experimental study on epinions. com community. In *AAAI*, volume 5, pages 121–126, 2005.

[39] Arya Mazumdar and Barna Saha. Query complexity of clustering with side information. In *Advances in Neural Information Processing Systems*, pages 4685–4696, 2017.

[40] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[41] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2008.

[42] Federico Monti, Michael Bronstein, and Xavier Bresson. Geometric matrix completion with recurrent multi-graph neural networks. In *NIPS*, pages 3700–3710, 2017.

[43] Elchanan Mossel and Jiaming Xu. Density evolution in the degree-correlated stochastic block model. *arXiv preprint arXiv:1509.03281*, 7, 2015.

[44] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *NIPS*, pages 2796–2804, 2013.

[45] Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. In *NIPS*, pages 2107–2115, 2015.

[46] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, pages 713–719, 2005.

[47] Hussein Saad and Aria Nosratinia. Community detection with side information: Exact recovery under the stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[48] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *ICML*, pages 880–887, 2008.

[49] Devavrat Shah and Christina Lee Yu. Reducing crowdsourcing to graphon estimation, statistically. *arXiv preprint arXiv:1703.08085*, 2017.

[50] Luo Si and Rong Jin. Flexible mixture model for collaborative filtering. In *ICML*, pages 704–711, 2003.

[51] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[52] Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *stat*, 1050:7, 2017.

[53] Rui Wu, Jiaming Xu, Rayadurgam Srikant, Laurent Massoulié, Marc Lelarge, and Bruce Hajek. Clustering and inference from pairwise comparisons. In *ACM SIGMETRICS Performance Evaluation Review*, volume 43, pages 449–450. ACM, 2015.

[54] Xiwang Yang, Yang Guo, and Yong Liu. Bayesian-inference-based recommendation in online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(4):642–651, 2013.

[55] Xiwang Yang, Harald Steck, Yang Guo, and Yong Liu. On top-k recommendation using social networks. In *RecSys*, pages 67–74. ACM, 2012.

[56] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. In *NIPS*, pages 4152–4160, 2016.

[57] Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014.

[58] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[59] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*, 2015.

[60] Huan Zhao, Quanming Yao, James T Kwok, and Dik Lun Lee. Collaborative filtering with social local models. In *ICDM*, pages 645–654, 2017.