When to Use Graph Side Information in Matrix Completion

Geewon Suh
EE, KAIST
Email: gwsuh91@kaist.ac.kr

Sangwoo Jeon
EE, KAIST
Email: sangw0@kaist.ac.kr

Changho Suh
EE, KAIST
Email: chsuh@kaist.ac.kr

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518134

***Abstract*—We consider a matrix completion problem that leverages graph as side information. One common approach in recently developed efficient algorithms is to take a two-step procedure: (i) clustering communities that form the basis of the graph structure; (ii) exploiting the estimated clusters to perform matrix completion together with iterative local refinement of clustering. A major limitation of the approach is that it achieves the information-theoretic limit on the number of observed matrix entries, promised by maximum likelihood estimation, only when a sufficient amount of graph side information is provided (the quantified measure is detailed later). The contribution of this work is to develop a computationally efficient algorithm that achieves the optimal sample complexity *for the entire regime of graph information*. The key idea is to make a careful selection for the information employed in the first clustering step, between two types of given information: graph & matrix ratings. Our experimental results conducted both on synthetic and real data confirm the superiority of our algorithm over the prior approaches in the scarce graph information regime.**

A full version of this paper is accessible at: https://sites.google.com/view/gwsuh/home/full-version_isit2021

I. Introduction

Recommender systems (RSs) aim to provide users with relevant items of their potential preference and interest. During the last decade, low-rank matrix completion, a prominent technique for operating RSs, have been extensively investigated and shown to be powerful in a wide variety of applications [1–8]. One challenge that arises in practice is the so-called *cold start problem*: high-quality recommendation is not ensured for new users and/or items. A natural way to address the problem is to exploit additional side information. Indeed, it has been demonstrated that the use of social networks such as Facebook's friendship graph [9] can improve the quality of recommendation [10–19].

Recent efforts have been made to provide information-theoretical insights into the gain due to graph side information [19–24]. In particular, the pioneering work [19] characterized the optimal sample complexity of matrix completion as a function of the *quality* of a given social graph. In the considered stochastic block model (SBM), the quality is quantified as $I_s := (\sqrt{\alpha}-\sqrt{\beta})^2$ where α (or β) indicates the edge probability between two users in the *same* (or *different*) clusters. This work has later been generalized to more practically relevant scenarios [20–22] with the aid of such (or similarly) quantified graph information.

One practical limitation of the prior works is that the computationally efficient algorithms proposed therein achieve the optimal sample complexity, only when the amount of graph information is sufficiently large. For instance, in the two-cluster binary matrix setting [19], the achievable regime w.r.t. I_s reads $I_s = \omega(\frac{1}{n})$ where n denotes the number of users in the given social graph. Efficient algorithms for the entire I_s regime have been out of reach.

Contributions: Our contribution lies in the development of a computationally efficient algorithm that ensures optimality *for the entire range of* I_s. As an initial effort, we focus on the simple two-cluster binary matrix setting considered in [19], although it can readily be extensible to other settings; for details, see Remark 2 in Section IV.

One key feature in the prior efficient algorithms [19–24] is that they take a well-known two-step procedure [5, 25–28] in which clustering is first done using a spectral method and then matrix ratings are estimated followed by iterative local refinement of clustering. We find that the sole use of a given graph in the first clustering step limits the applicability to the scare graph information regime. Inspired by this, we develop a switching-geared clustering strategy which selects employed information between graph and matrix ratings, depending on the amount of graph side information; see Algorithm 1 for details on the threshold for switching. Employing perturbation techniques in random matrix theory [29, 30], we show that our algorithm indeed achieves the optimal sample complexity for the entire I_s regime. We also conduct experiments both on synthetic and real datasets to demonstrate the superior performance over other approaches [15, 19, 31–33] including the ones that rely solely upon graph information in the first clustering stage.

Related works: In addition to [19], graph-assisted matrix completion has been explored for various settings [20–24]. Yoon et al. [20] considered a more-than-two-cluster model. Elmahdy et al. [21] considered a more practically relevant scenario in which clusters exhibit a hierarchical structure. Jo et al. [22] introduced a generalized rating model in which each matrix entry can take an arbitrary discrete value. Zhang et al. [23, 24] explored a richer scenario wherein both social and item similarity graphs are available as side information. While all of the works characterized optimal sample complexities for the considered models together with the development of

efficient algorithms, their algorithms are far from optimality when the amount of associated graph information is not sufficiently large. In contrast, our algorithm ensures optimality for the entire graph information regime under the two-cluster setting; see Remark 2 for generalization.

As mentioned earlier, the key idea of our algorithm is centered on the first clustering step which builds upon prominent graph-based clustering [25, 26, 29, 34] as well as other clustering methods [35–38]. Our switching mechanism in clustering also employs matrix-rating-based clustering aided by singular value decomposition (SVD) [35, 36]. We find that several techniques [30, 39–42] intended for the SVD-based approaches help us to analyze our algorithm in our setting. Technical contributions are reflected in Lemmas 2 and 3.

Notation: We define that a clustering algorithm guarantees "weak clustering" or "weak recovery" if an algorithm allows for a vanishing fraction of misclassified users as the number of users tends to infinity.

II. Problem Formulation

Setting: As an initial effort, we focus on the simple setting as in [19], in which a rating matrix consists of nm entries for n users (rows) and m items (columns). Below is a list of assumptions made for theoretical guarantees of our algorithm (Theorem 3), but not for the algorithm itself (Algorithm 1). Each user rates items either as 1 or -1 (for instance, "like" or "dislike"). We assign 0 for unrated items. Assume that there are two equal-sized clusters of users, say A and B, and the users from the same group share the same rating vector. Let $v_A, v_B \in \{-1,1\}^{1\times m}$ be the rating vectors of cluster A and B, respectively. Let $M \in \{-1,1\}^{n\times m}$ be a rating matrix where the ith row corresponds to the rating vector of user i. Let $\delta := \frac{1}{m} d_H(v_A, v_B)$ be the normalized Hamming distance between v_A and v_B, and $M^{(\delta)}$ be the collection of rating matrices such that the normalized Hamming distance of two rating vectors does not exceed δ.

Problem of interest: Our goal is to recover a rating matrix $M \in \mathcal{M}^{(\delta)}$ given two types of information. The first is a partially observed matrix $Y \in \{-1,0,1\}^{n\times m}$. We denote by Ω the set of observed entries of Y: $\Omega = \{(i,j) \in [n]\times[m] : Y_{ij} \neq 0\}$. We assume that each element of M is observed with probability $p \in [0,1]$, independently from others, and its observation can be flipped with probability $\theta \in [0, \frac{1}{2})$: $Y_{ij} \sim \text{Bern}(p) \cdot (1 - 2\text{Bern}(\theta)) \times M_{ij}$. The second is social graph $\mathcal{G} = ([n], E)$ where E denotes the set of edges, each capturing social connection. The set $[n]$ of vertices is partitioned into two disjoint clusters. We assume that the graph follows the SBM with two types of edge probabilities: α (or β) for intra-cluster (or cross-cluster) users. We focus on realistic scenarios where the same cluster users are more likely to be connected: $\alpha \geq \beta$.

Performance metric: Let $\psi(Y, \mathcal{G}) \in \mathbb{R}^{n\times m}$ be the estimator of a rating matrix. We use the worst-case error probability $P_e^{(\delta)}$ as a performance metric. The worst-case ground truth matrix is chosen subject to the normalized Hamming distance δ and the associated error probability reads: $P_e^{(\delta)}(\psi) := \max_{M\in\mathcal{M}^{(\delta)}} \mathbb{P}[\psi(Y,\mathcal{G}) \neq M]$. We intend to develop an efficient estimator ψ that satisfies $P_e^{(\delta)} \to 0$ as $n \to \infty$ for any p larger than p^*. Here p^* denotes the optimal sample probability: (i) if $p \geq p^*$, $P_e^{(\delta)} \to 0$ as $n \to \infty$ for some estimator ψ; (ii) if $p < p^*$, $P_e^{(\delta)} \not\to 0$ as $n \to \infty$ for any ψ.

III. Main Results

Let us start by recalling the optimal sample probability p^*, characterized in [19]. Let $I_s := (\sqrt{\alpha} - \sqrt{\beta})^2$ be a quantified measure for the quality of social graph; the higher, the easier to cluster and hence the more graph information. As in [19], we make the same assumption on n and m that turns out to ease the proof via large deviation theories: $m = \omega(\log n)$ and $\log m = o(n)$. This assumption is also practically relevant as it rules out highly asymmetric matrices.

Theorem 1 (Optimal sample probability [19]): Let

$$p^*(I_s) = p^* := \frac{1}{(\sqrt{1-\theta} - \sqrt{\theta})^2} \max\left\{ \frac{\log n - \frac{n}{2} I_s}{\delta m}, \frac{2\log m}{n} \right\}.$$

Fix $\epsilon > 0$. If $p > (1+\epsilon)p^*$, $P_e^{(\delta)}(\psi) \to 0$ as $n \to \infty$ for some sequence of estimator ψ. Conversely, if $p < (1-\epsilon)p^*$, $P_e^{(\delta)}(\psi) \not\to 0$ as $n \to \infty$ for any ψ.

We often use a simpler notation p^* for $p^*(I_s)$. Ahn et al. [19] also developed an efficient algorithm that achieves p^* for a certain range of I_s, formally stated below.

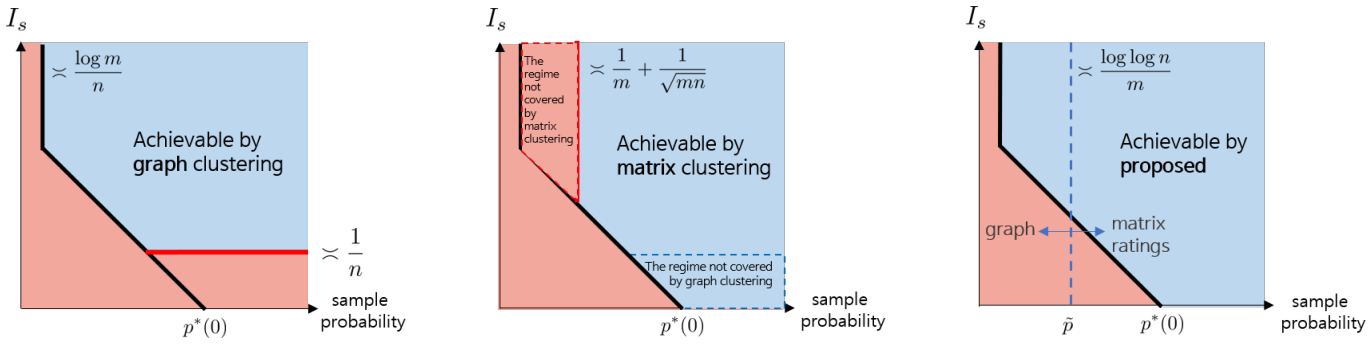
Theorem 2 (Theoretical guarantees of [19]'s algorithm): Suppose that $I_s = \omega\left(\frac{1}{n}\right)$ and p respects the sufficient condition in Theorem 1. Then, the algorithm in [19] exactly recovers M with high probability as n and m tend to infinity: $\mathbb{P}(\hat{M} = M) = 1 - o(1)$ where $\hat{M} := \psi(Y, \mathcal{G})$.

The existence of an optimal efficient algorithm guaranteed for the entire I_s regime has been unknown. We develop an efficient universal algorithm that promises p^* for the entire range of I_s.

Theorem 3 (Theoretical guarantees of our universal algorithm): Suppose that p respects the sufficient condition in Theorem 1. Then, our computationally-efficient algorithm described in Algorithm 1 exactly recovers M with high probability as n and m tend to infinity.

This implies that there is no information-computation gap for any regime of I_s. Even when $I_s = O\left(\frac{1}{n}\right)$, there exists an optimal efficient algorithm. In fact, the main reason that the algorithm in [19] offers the limited achievable regime is that it relies solely upon graph information for clustering in the first step. This motivates us to develop a switching-geared clustering strategy that properly chooses employed information for clustering between graph and matrix ratings. It turns out this leads to universal optimality.

Remark 1 (Comparison to other efficient algorithms): Fig. 1 illustrates achievable regimes (shaded in blue color) of (p, I_s) promised by an employed algorithm. The red shaded region indicates the non-achievable regime. The black bold line indicates the boundary dictated by the optimal sample probability $p^* = p^*(I_s)$. Here $p^*(0)$ denotes the case $I_s = 0$. Fig. 1(a) refers to the achievable regime for the algorithm in



(a) Graph clustering (b) Matrix clustering (c) Proposed

Fig. 1: Achievable regimes (shaded in blue color) due to: (a) graph-clustering approach (use only graph in the first clustering step); (b) matrix-rating-clustering approach (use only matrix ratings); (c) our proposed approach (use graph or matrix ratings depending on the amount of graph information). With proper switching of the employed information for clustering, our algorithm achieves the entire achievable regime promised by MLE.

[19] that employs only graph information in the first clustering step. As implied by Theorem 2, achievability is guaranteed only when $I_s = \omega\left(\frac{1}{n}\right)$; see the red thick line on the right bottom. Fig. 1(b) refers to the case in which we employ only matrix ratings for clustering in the first stage (explicit details on this are left in Algorithm 1). In this case, we could show achievability when I_s is small: $I_s = O\left(\frac{1}{n}\right)$. See a dotted box on the right bottom: the small I_s regime becomes blue. On the other hand, we find that the high I_s regime can be translated to $p = O\left(\frac{1}{m} + \frac{1}{\sqrt{nm}}\right)$ along the optimal boundary indicated by the black bold line. We also find that in such p regime above (guided by the red thick line), achievability is not guaranteed; see the red-colored regime on the left top. Details on these arguments as well as on how to translate the high I_s regime into the above p regime will be provided while describing the proof of Theorem 3 in Section IV; see Lemma 1 in particular. On the other hand, by switching clustering methods, our algorithm ensures achievability for the entire (p, I_s) regime promised by MLE. See Fig. 1(c). ■

IV. Proof of Theorem 3

A. Algorithm Description

As hinted earlier, our algorithm builds upon a well-known two-stage procedure [5, 25–28]. For illustrative purpose, as in [19, 20, 22, 24], we split the second stage into two steps, thereby having three steps in total. Step 1 is the major step that bears the key idea. In Step 1, we intend to identify two clusters based on graph and matrix ratings. The 2nd and 3rd steps are the standard steps employed in [19]. In Step 2, we use the estimated clusters to recover two rating vectors. In Step 3, we do iterative refinement of clustering based on the estimated vectors. The details are described below; also see Algorithm 1.

Step 1 (Initial clustering via a switching mechanism): Inspired by the observations in Fig. 1(a) and (b), we develop a switching mechanism. Notice that the regime $I_s = O(\frac{1}{n})$ is not achievable with graph-based clustering [43], yet being achievable with matrix-rating-based clustering; and vice versa for the other regime $I_s = \omega(\frac{1}{n})$. This naturally motivates a unified approach that makes a proper selection between the two clustering methods depending on the amount of I_s. However, there is a challenge in implementing this natural idea. The challenge is that the estimation of I_s is not that simple as it requires the knowledge of (α, β), which are difficult to estimate without ground-truth clusters.

Hence, we propose an alternative that relies upon another easily computable information yet which plays the same role. The idea is inspired by the optimal boundary indicated by the black bold line in Fig. 1. Notice along the optimal boundary that the high I_s regime corresponds to the small $p^*(I_s)$ regime and vice versa. This suggests that p can play the same role for switching. For a small p (a large I_s), we apply graph-based clustering. For a large p (a small I_s), we perform matrix-rating-based clustering. Here a practical benefit from the use of p is that it is easy to estimate using observed ratings, and also its MLE is accurate in the large (n, m) regime of practical interest: $\hat{p} := \frac{|\Omega|}{nm}$ converges to p in the limit of n and m.

Now a key question is: what is a proper switching threshold, say $\tilde{p}$? To answer this question, we intend to identify the range of p in which the optimal sample probability is achievable with matrix-rating-based clustering a.k.a. matrix clustering. It turns out that matrix clustering leads to optimality as long as $p = \omega\left(\frac{1}{m} + \frac{1}{\sqrt{nm}}\right)$. This is proved in Lemma 2 in Section IV-B. This motivates the choice of $\tilde{p}$ as the one that is order-wise greater than $\frac{1}{m} + \frac{1}{\sqrt{nm}}$. Here we made one particular choice like $\tilde{p} = \frac{\log\log n}{m}$. Also one can verify that $p > \tilde{p}$ covers the low $I_s = O\left(\frac{1}{n}\right)$ regime, as hinted in Fig. 1(c), and hence matrix clustering ensures optimality for the low I_s regime not covered by graph clustering.

Matrix-rating-based clustering: If the empirical estimate $\hat{p}$ is beyond the threshold $\tilde{p} = \frac{\log\log n}{m}$, we perform matrix-based clustering. We first perform SVD w.r.t. the observed matrix Y: $Y = U\Sigma V^T$. We then generate an n-by-2 matrix U_Y such that it takes only the two leading columns of U. Next we

apply the famous k-means algorithm [44] w.r.t. U_Y to yield the estimated clusters, say $A^{(0)}$ and $B^{(0)}$. For implementation, see lines 3–6 in Algorithm 1. We will later show that this procedure guarantees weak clustering, thereby ensuring matrix completion together with Steps 2 and 3. See Lemma 2 in Section IV.B for the weak clustering guarantee.

Graph-based clustering: If $\hat{p} < \tilde{p}$, we employ graph clustering [25] to yield the estimated clusters. See lines 7 and 8. It has been shown in [25] that graph clustering in the considered regime ensures weak clustering.

Step 2 (Recovery of rating vectors): This step is exactly the same as that in [19]. Based on $A^{(0)}$ and $B^{(0)}$, we estimate (v_A, v_B) via majority voting. For each item j, we set the jth entry of $\hat{v}_A$ (estimate), as the majority among the observed ratings in the jth columns w.r.t. the rows governed by $A^{(0)}$. Similarly we estimate $\hat{v}_B$ using $B^{(0)}$. See lines 10–14 for implementation. It was shown in [19] that this majority voting ensures exact recovery of rating vectors.

Step 3 (Local refinement of clustering): Again this step is identical to that in [19]. We do iterative local refinement of clustering $(A^{(0)}, B^{(0)})$ using localized MLE in which the likelihood is computed based on the estimated clusters in the prior iteration. See lines 19–27 for implementation. We apply $T = O(\log n)$ iterations, as it is shown that such number guarantees exact clustering [19].

Remark 2 (Generalization): One can extend the proposed switching mechanism to other settings considered in [20–24]. The idea is to draw similar plots as in Fig. 1(a) and (b), and then identify a proper threshold $\tilde{p}$ for switching with the help of some follow-up analysis of the associated matrix clustering. The detailed analysis is out of the scope of this work. ■

B. Proof Outline

Due to space limit, we provide only the sketch of the proof, while leaving the complete proof in the full version. The proof consists of two parts. The first is to show that the achievability proof boils down to the proof of weak recovery of matrix clustering when $p = \omega(\frac{1}{m} + \frac{1}{\sqrt{nm}})$. This will be proved in Lemma 1. The second is to prove the weak recovery of matrix clustering for the focused regime (Lemma 2). Most of the proofs of lemmas are left in the complete version.

To prove the first part, we introduce two regimes: Regime 1 $:= \{(p, I_s) : p \geq p^*, p < \tilde{p}\}$, Regime 2 $:= \{(p, I_s) : p \geq p^*, p \geq \tilde{p}\}$. We claim that: (i) in Regime 1, $p < \tilde{p}$ implies $I_s = \omega\left(\frac{1}{n}\right)$; (ii) in Regime 2, $p \geq \tilde{p}$ covers $I_s = O(\frac{1}{n})$. This claim is proved in Lemma 1.

Lemma 1: If $p \geq p^*$ and $p < \tilde{p}$, then $I_s = \omega\left(\frac{1}{n}\right)$. Also if $p \geq p^*$ and $I_s = O\left(\frac{1}{n}\right)$, then $p \geq \tilde{p}$.

The second part stated in Lemma 1 implies that $p \geq \tilde{p}$ covers the remaining regime $I_s = O\left(\frac{1}{n}\right)$ (not covered by graph clustering) as long as $p \geq p^*$. Hence, for Regime 2, it suffices to prove weak recovery of matrix clustering. The proof of this is in Lemma 2.

Lemma 2: If $p \geq p^*$ and $p = \omega\left(\frac{1}{m} + \frac{1}{\sqrt{nm}}\right)$, matrix-rating-based clustering in Step 1 guarantees weak recovery.

Algorithm 1: Proposed Algorithm

Input : Observed rating matrix $Y \in \{-1, 0, 1\}^{n \times m}$
Graph $\mathcal{G} = ([n], E)$
The number of iteration for refinement T
Output: Estimate of a rating matrix $\hat{M} \in \{-1, 1\}^{n \times m}$

1 $\hat{p} \leftarrow |\Omega|/nm$;
2 **Step 1 (Initial clustering via a switching mechanism)**
3 **if** $\hat{p} > \tilde{p} = \frac{\log\log n}{m}$ **then**
4 $U\Sigma V^T \leftarrow$ singular value decomposition of Y;
5 $U_Y \leftarrow$ two leading columns of U;
6 Apply the k-means clustering w.r.t. U_Y to obtain an initial estimate for clustering: $(A^{(0)}, B^{(0)})$;
7 **else**
8 Apply graph-based clustering w.r.t. G to obtain an initial estimate for clustering: $(A^{(0)}, B^{(0)})$;
9 **end**
10 **Step 2 (Recovery of rating vectors)**
11 **for** *item* $j = 1$ *to* m **do**
12 $(\hat{v}_A)_j \leftarrow \text{sign}(\sum_{i \in A^{(0)}} Y_{ij})$;
13 $(\hat{v}_B)_j \leftarrow \text{sign}(\sum_{i \in B^{(0)}} Y_{ij})$;
14 **end**
15 **Step 3 (Local refinement of clustering)**
16 $\hat{\alpha} \leftarrow \frac{1}{\binom{|A^{(0)}|}{2} + \binom{|B^{(0)}|}{2}} |\{\{i_1, i_2\} \in E : i_1, i_2 \in A^{(0)} \text{ or } i_1, i_2 \in B^{(0)}\}|$;
17 $\hat{\beta} \leftarrow \frac{1}{|A^0||B^{(0)}|} |\{\{i_1, i_2\} \in E : i_1 \in A^{(0)}, i_2 \in B^{(0)}\}|$;
18 $\hat{\theta} \leftarrow |\{(i,j) \in [n] \times [m] : Y_{ij} \neq 0, Y_{ij} \neq (\hat{v}_A)_j, i \in A^{(0)} \text{ or } Y_{ij} \neq (\hat{v}_B)_j, i \in B^{(0)}\}|/|\Omega|$;
19 **for** *iteration* $t = 1$ *to* T **do**
20 $A^{(t)}, B^{(t)} \leftarrow \emptyset$;
21 **for** *user* $i = 1$ *to* n **do**
22 $L_A(i) \leftarrow \log\left(\frac{(1-\hat{\beta})\hat{\alpha}}{(1-\hat{\alpha})\hat{\beta}}\right) e(\{i\}, A^{(t-1)}) + \log\left(\frac{1-\hat{\theta}}{\hat{\theta}}\right) \sum_{j \in [m]} \mathbb{1}(Y_{ij} = (\hat{v}_A)_j)$;
23 $L_B(i) \leftarrow \log\left(\frac{(1-\hat{\beta})\hat{\alpha}}{(1-\hat{\alpha})\hat{\beta}}\right) e(\{i\}, B^{(t-1)}) + \log\left(\frac{1-\hat{\theta}}{\hat{\theta}}\right) \sum_{j \in [m]} \mathbb{1}(Y_{ij} = (\hat{v}_B)_j)$;
24 **if** $L_A(i) > L_B(i)$ **then** $A^{(t)} \leftarrow A^{(t)} \cup \{i\}$;
25 **else** $B^{(t)} \leftarrow B^{(t)} \cup \{i\}$;
26 **end**
27 **end**
28 **for** *user* $i = 1$ *to* n **do**
29 **if** $i \in A^{(T)}$ **then** $\hat{M}_i \leftarrow \hat{v}_A$ **else** $\hat{M}_i \leftarrow \hat{v}_B$;
30 **end**
31 **Return** $\hat{M}$

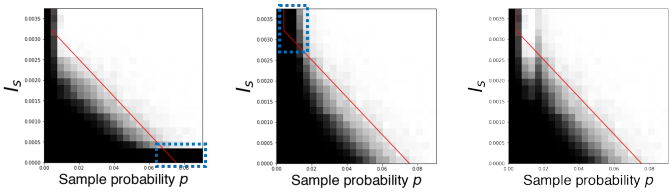
Proof: Let us first introduce some notations. Recall from Algorithm 1 that U_Y consists of the two leading left singular vectors of Y. Let $\sqrt{\frac{2}{n}}C_A, \sqrt{\frac{2}{n}}C_B \in \mathbb{R}^{1 \times 2}$ denote the centers among the points (each referring to a certain row in U_Y) that correspond to A and B, respectively. Let U_G be an n-by-2 matrix such that the ith row of $U_G = \sqrt{\frac{2}{n}}C_A$ if $i \in A$; $\sqrt{\frac{2}{n}}C_B$ otherwise. Here what we can show is that by using the approximate k-means error bound in [29] (see Lemma 5.3 therein), the fraction of misclustered users is bounded by $\|U_Y - U_G\|_F^2$ up to a constant factor. For completeness, we also leave the detailed proof in the full version. Hence, it suffices to show $\|U_Y - U_G\|_F^2 \to 0$ for proving weak recovery. This proof is done in Lemma 3. This completes the proof of Lemma 2. ■

Lemma 3: If $p = \omega\left(\frac{1}{m} + \frac{1}{\sqrt{nm}}\right)$, $\|U_Y - U_G\|_F^2 \to 0$ with high probability as $n, m \to \infty$.

Remark 3 (Technical novelty): One major technical contribution is reflected in Lemma 3. The key step in the proof is to show that U_Y and the two leading ground-truth singular vectors are very similar. To this end, we employ perturbation bounding technique for singular subspaces in [29]. We then derive an upper bound of $\sin\Theta$ distance between U_Y and the ground-truth singular vectors as a function of the variance of Y_{ij}'s. Lastly we prove that the upper bound converges to 0. ■

V. Experiments

We provide Monte Carlo experiments to corroborate our main results of Theorem 3.



(a) Graph clustering (b) Matrix clustering (c) Proposed

Fig. 2: Achievable regimes of (p, I_s) due to: (a) graph-based clustering; (b) matrix-rating-based clustering; (c) our proposed algorithm. Here the brightness indicates the level of the empirical success rate; the brighter, the higher.

Synthetic data: Synthetic data is generated as per the model described in Section II. Here, we consider a setting where $\theta = 0.1$, $n = 1000$, $m = 100$ and $\delta = 0.5$. In Fig. 2, we evaluate the performance of three algorithms via the empirical success rate for a rage of (p, I_s). The empirical success rate is visualized by the brightness level; the brighter, the higher. The red line indicates the sharp boundary indicated by the optimal sample probability $p^* = p^*(I_s)$. Fig. 2(a) shows performance of graph-clustering-based approach. As illustrated in Fig. 1(a), we also observe the failure of recovery in the low I_s regime. See the lower right corner of Fig. 2(a). On the other hand, in Fig. 2(b) w.r.t. matrix-rating-based approach, recovery is mostly successful in the low p regime. As illustrated in Fig. 1(b), however, it suffers from performance degradation in the high I_s regime; see the upper left corner. By switching between two clustering methods depending on $\hat{p}$ (MLE of p), our algorithm exhibits an improved performance for the two focused regimes; see Fig. 2(c).

We also plot the performance for the low I_s regime (the scarce graph information regime, highlighted in Theorem 3). To this end, we consider a setting where $\theta = 0.1$, $\delta = 0.5$, $I_s = \frac{1}{n}$, and we vary n and m while preserving $n/m = 5$. In Fig. 3, we plot the empirical success rate as a function of normalized sample complexity $\frac{p}{p^*}$. We observe that the empirical success rate gets closer to 1 as soon as p exceeds p^*, and the transition becomes sharper with an increase in n. *Real data*: As in [19–22, 24], we consider a semi-real data setting in which a social graph is real while rating vectors are synthetically generated as per our considered model.

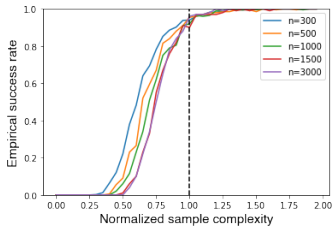
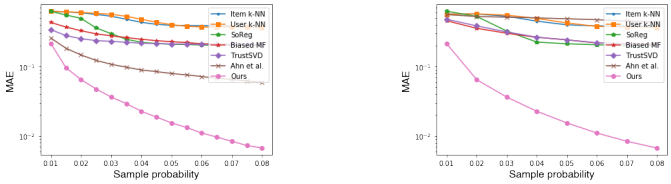


Fig. 3: The empirical success rate of the proposed algorithm as a function of p/p^* when $n = 5m$ and $I_s = \frac{1}{n}$ $(I_s = O\left(\frac{1}{n}\right))$.



(a) (b)

Fig. 4: Comparison of MAEs evaluated for various recommendation algorithms using political blog network [45]: (a) $\hat{I}_s = 4.67 \times 10^{-3}$ (high I_s regime); (b) $\hat{I}_s = 8.07 \times 10^{-4}$ (low I_s regime).

As a real graph, we employ political blog network having the ground-truth clusters and $n = 1222$ [45]. We consider two settings: (i) the original blog network; (ii) a sparse network generated via randomly subsampling 5% of edges from the original. We synthesize ratings to generate a rating matrix with $(n, m) = (1222, 500)$. We use mean absolute error (MAE), $\sum_{(i,j)\in[n]\times[m]} |\hat{M}_{ij} - M_{ij}|/nm$, as a metric to compare ours with various recommendation algorithms exploiting graph side information: (i) item k-nearest neighbor (k-NN) [31]; (ii) user k-NN [31]; (iii) matrix factorization and social regularization (SoReg) [15]; (iv) biased matrix factorization (Biased MF) [32]; (v) TrustSVD [33]; and (vi) Ahn et al. [19] based solely on graph clustering. As shown in Fig. 4, our algorithm outperforms all of the baselines. In particular, the gain is more significant in the interested small I_s regime $(\hat{I}_s = O\left(\frac{1}{n}\right))$.

VI. Discussion

We develop an efficient algorithm that achieves the optimal sample complexity for the entire range of I_s. The key idea is to take a switching-geared clustering strategy which carefully selects employed information for clustering between graph and matrix ratings, depending on the amount of graph side information.

To simplify our algorithm, one can consider an automatic switching mechanism by minimizing some combination of graph clustering error and matrix clustering error. If we can prove the theoretical guarantee of this simplified algorithm, it would strengthen our result. Another future work of interest is to extend to more practically relevant settings, via relaxing the assumptions made in our considered model, as mentioned in Remark 2.

REFERENCES

[1] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.

[2] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[3] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[4] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[5] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[7] K. Lee and Y. Bresler, "Admira: Atomic decomposition for minimum rank approximation," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4402–4416, 2010.

[8] L. T. Nguyen, J. Kim, and B. Shim, "Low-rank matrix completion: A contemporary survey," *IEEE Access*, vol. 7, pp. 94 215–94 237, 2019.

[9] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012.

[10] C. C. Aggarwal, J. L. Wolf, K.-L. Wu, and P. S. Yu, "Horting hatches an egg: A new graph-theoretic approach to collaborative filtering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 201–212.

[11] P. Massa and P. Avesani, "Trust-aware collaborative filtering for recommender systems," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2004, pp. 492–508.

[12] J. Golbeck, J. Hendler *et al.*, "Filmtrust: Movie recommendations using trust in web-based social networks," in *Proceedings of the IEEE Consumer Communications and Networking Conference*, vol. 96, no. 1. Citeseer, 2006, pp. 282–286.

[13] M. Jamali and M. Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 397–406.

[14] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2010.

[15] 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.

[16] X. Yang, Y. Guo, and Y. Liu, "Bayesian-inference-based recommendation in online social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 4, pp. 642–651, 2012.

[17] J. Tang, X. Hu, and H. Liu, "Social recommendation: a review," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1113–1133, 2013.

[18] F. Monti, M. Bronstein, and X. Bresson, "Geometric matrix completion with recurrent multi-graph neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 3697–3707.

[19] K. Ahn, K. Lee, H. Cha, and C. Suh, "Binary rating estimation with graph side information," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 4272–4283.

[20] 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.

[21] A. Elmahdy, J. Ahn, C. Suh, and S. Mohajer, "Matrix completion with hierarchical graph side information," in *34th Conference on Neural Information Processing Systems, NeurIPS 2020*, 2020.

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

[23] Q. Zhang, V. Y. F. Tan, and C. Suh, "Community detection and matrix completion with social and item similarity graphs," *to appear in the IEEE Transactions on Signal Processing*, 2021.

[24] Q. Zhang, G. Suh, C. Suh, and V. Y. Tan, "Mc2g: An efficient algorithm for matrix completion with social and item similarity graphs," *arXiv preprint arXiv:2006.04373*, 2020.

[25] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block models," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1980–2024, 2017.

[26] 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*. IEEE, 2015, pp. 670–688.

[27] Y. Chen and C. Suh, "Spectral mle: Top-k rank aggregation from pairwise comparisons," in *International Conference on Machine Learning*. PMLR, 2015, pp. 371–380.

[28] Y. Chen, G. Kamath, C. Suh, and D. Tse, "Community recovery in graphs with locality," in *International Conference on Machine Learning*. PMLR, 2016, pp. 689–698.

[29] J. Lei and A. Rinaldo, "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, vol. 43, no. 1, p. 215–237, Feb 2015.

[30] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *The Annals of Statistics*, vol. 46, no. 1, pp. 60–89, Feb 2018.

[31] R. Pan, P. Dolog, and G. Xu, "Knn-based clustering for improving social recommender systems," in *International workshop on agents and data mining interaction*. Springer, 2012, pp. 115–125.

[32] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 426–434.

[33] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[34] P. Chin, A. Rao, and V. Vu, "Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery," in *Conference on Learning Theory*, 2015, pp. 391–423.

[35] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for k-means clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, 2014.

[36] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering," *SIAM Journal on Computing*, vol. 49, no. 3, pp. 601–657, 2020.

[37] H. Ashtiani, S. Kushagra, and S. Ben-David, "Clustering with same-cluster queries," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 3224–3232.

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

[39] M. Azizyan, A. Singh, and L. Wasserman, "Minimax theory for high-dimensional gaussian mixtures with sparse mean separation," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2139–2147, 2013.

[40] J. Jin and W. Wang, "Influential features pca for high dimensional clustering," *The Annals of Statistics*, pp. 2323–2359, 2016.

[41] J. Jin, Z. T. Ke, W. Wang *et al.*, "Phase transitions for high dimensional clustering and related problems," *The Annals of Statistics*, vol. 45, no. 5, pp. 2151–2189, 2017.

[42] V. Vu, "A simple svd algorithm for finding hidden partitions," *Combinatorics, Probability & Computing*, vol. 27, no. 1, p. 124, 2018.

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

[44] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[45] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, 2005, pp. 36–43.