The Optimal Sample Complexity of Matrix Completion with Hierarchical Similarity Graphs

Adel Elmahdy
ECE, University of Minnesota
Email: adel@umn.edu

Junhyung Ahn
EE, KAIST
Email: tonyahn96@kaist.ac.kr

Soheil Mohajer
ECE, University of Minnesota
Email: soheil@umn.edu

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

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834449

***Abstract*—We study a matrix completion problem that leverages a hierarchical structure of social similarity graphs as side information in the context of recommender systems. We assume that users are categorized into clusters, each of which comprises sub-clusters (or what we call "groups"). We consider a low-rank matrix model for the rating matrix, and a hierarchical stochastic block model that well respects practically-relevant social graphs. Under this setting, we characterize the information-theoretic limit on the number of observed matrix entries (i.e., optimal sample complexity) as a function of the quality of graph side information (to be detailed) by proving sharp upper and lower bounds on the sample complexity. Furthermore, we develop a matrix completion algorithm and empirically demonstrate via extensive experiments that the proposed algorithm achieves the optimal sample complexity.**

I. Introduction

Personalized recommender systems have emerged in a wide range of Web applications to predict the preferences of its users and provide them with new and relevant items based on scarce data about the users and/or items [1]. Inspired by the Netflix challenge, a well-known technique for predicting the missing ratings in collaborative filtering frameworks is low-rank matrix completion. Given partial observation of a matrix of users by items, the goal is to develop an algorithm to accurately predict the values of the missing ratings. One of the prime challenges of collaborative filtering systems that rely on user-item interactions is the "cold start problem" in which high-quality recommendations are not feasible for new users/items that bear little or no information. A prominent technique to overcome this problem is to incorporate the community information into the framework of recommender systems in order to enhance the recommendation quality.

Numerous research works have explored the idea of leveraging the information inferred from social graphs to enhance the performance of recommender systems from an algorithmic perspective [1]–[20]. Recently, [21]–[24] have investigated the problem of interest from an information-theoretic perspective. However, they impose a number of strict assumptions on the system model such as the users of the same cluster have same ratings over all items, and hence each cluster is represented by a rank-one matrix. This limits the practicality of the proposed models for real-world data. In this work, we relax this assumption and study a more generalized framework in which each cluster is represented by a rank-r matrix. In particular, we consider a matrix completion problem where the users are categorized into c clusters, each of which comprises g sub-clusters, or what we call "groups", producing a hierarchical structure in which the features of different groups within a cluster are broadly similar to each other; however, they are different from the features of the groups in other clusters.

The main contributions of this paper are summarized as follows. We characterize an information-theoretic threshold for reliable matrix recovery as a function of the quantified quality of the considered hierarchical graph side information by establishing matching upper and lower bounds on the sample complexity. To the best of our knowledge, this is the first work to provide this characterization for any finite field size, and any number of clusters and groups. We show that the proposed algorithm, which leverages the hierarchical graph structure, yields a substantial gain in sample complexity, compared to a simple variant of [21], [22] that does not leverage the relational structure across rating vectors of groups. We also reveal that when the graph information is rich enough to perfectly retrieve the structures of clusters and groups, the optimal sample complexity increases linearly as the number of clusters increases. Otherwise, the optimal sample complexity remains almost constant, even though the number of groups in a cluster increases. Furthermore, we develop a matrix completion algorithm that starts with hierarchical graph clustering, which produces an exact recovery of clusters, but an almost exact recovery of groups. Then, the rating vectors are estimated followed by iterative local refinement of groups. We conduct extensive experiments to demonstrate that the optimal sample complexity is achieved by the proposed algorithm.

II. Problem Formulation

Consider a rating matrix $X \in \mathbb{F}_q^{n\times m}$, where n denotes the number of users and m denotes the number of items. The ratings of the rth user over m items forms the rth row of X for $r \in [n]$. However, the rating matrix is incomplete, in the sense that some entries might be missing. The user similarity graphs (e.g., social graphs) are leveraged as side-information to enhance the quality of the matrix completion. More specifically, we consider a hierarchical similarity graph over the users that consists of c disjoint *clusters*, and each cluster comprises g disjoint *groups*. For the sake of tractable

The work of A. Elmahdy and S. Mohajer is supported in part by the National Science Foundation under Grants CCF-1749981.

A. Elmahdy and J. Ahn contributed equally to this work.

mathematical analysis, we assume equal-sized clusters and groups. The theoretical guarantees, however, hold as long as the group sizes are order-wise same (See Section III). According to the social homophily theory [25], users within the same community (that is, those who are more likely to be connected in the social graph) are more likely to share similar preferences over items. This results in a low rank structure of the rating matrix since the rows of the rating matrix associated with such users are likely to be similar [26]. To capture this crucial fact in our model, we make the following assumptions: (i) the rating vectors of the users who belong to the same group are equal, and hence there are gc distinct rating vectors in total; (ii) the rating vectors of the groups of a given cluster are different, yet intimately-related to each other through a linear subspace of r basis vectors for some integer $r \leq g$ [27], [28]. Let $v_i^{(x)}$ denotes the rating vector of the users in cluster x and group i for $x \in [c]$ and $i \in [g]$. Let $R^{(x)} \in \mathbb{F}_q^{g \times m}$ denote a matrix whose rows are the rating vectors of the groups in cluster x for $x \in [c]$. The set of g rows of $R^{(x)}$ (that is, the set of g rating vectors of the groups in cluster x) is spanned by any subset of r rows of $R^{(x)}$. Let X_0 denote the ground truth rating matrix. Each instance of the problem corresponds to a rating matrix X_0, which can be represented by a set of rating vectors $\mathcal{V}_0 = \{u_i^{(x)} : x \in [c], i \in [g]\}$ and a user partitioning $\mathcal{Z}_0$. Formally, $\mathcal{Z}_0$ is a family of subsets of $[n]$ that partitions the set of all users $[n]$ into c clusters and g groups (per cluster).

The main goal is to find the best estimate of X_0 with the knowledge of two types of observations. The first type is a partial and noisy observation Y of X_0. For every $r \in [n]$ and $t \in [m]$, let $Y(r,t) \in \mathbb{F}_q \cup \{*\}$, where $*$ denotes no observation. Let the set of observed entries of X_0 be denoted by $\Omega = \{(r,t) \in [n] \times [m] : Y(r,t) \neq *\}$. The partial observation is modeled by assuming that each entry of X_0 is observed with probability $p \in [0,1]$, independently from others. Moreover, the potential noise in the observation is modeled by a random uniform noise distribution; that is, the noise is not adversarial (i.e., not deterministic). We assume that each observed entry $X_0(r,t)$, for $(r,t) \in \Omega$, can possibly be flipped to any element of the set $\{0,1,\ldots,q-1\} \backslash X_0(r,t)$ with a uniform probability of $\theta/(q-1)$ for $\theta \in [0,(q-1)/q)$. The second type of observation is user similarity graph $\mathcal{G} = ([n], \mathcal{E})$. A vertex represents a user, and an edge captures a social connection between two users. The set $[n]$ of vertices is partitioned into c disjoint clusters, each of which has n/c users. Each cluster is further partitioned into g disjoint groups, each of which has $n/(cg)$ users. The user similarity graph is generated according to the hierarchical stochastic block model (HSBM) [29], [30], which is a generative model for random graphs exhibiting hierarchical cluster behavior. In this model, each two nodes in the graph are connected by an edge, independent of all other nodes, such that there is an edge between two users in the *same group* within a cluster with probability $\widetilde{\alpha}$; there is an edge between two users in *different groups* but within the *same cluster* with probability $\widetilde{\beta}$; and there is an edge between two users in *different clusters* with probability $\widetilde{\gamma}$. We assume the edge probabilities scale with the problem size, i.e., $\widetilde{\alpha} = \alpha \frac{\log n}{n}$, $\widetilde{\beta} = \beta \frac{\log n}{n}$ and $\widetilde{\gamma} = \gamma \frac{\log n}{n}$, where α, β and γ are positive real numbers such that $\alpha \geq \beta \geq \gamma$. Note that the considered edge probabilities guarantee the disappearance of isolated vertices (i.e., vertices of degree zero) in the user similarity graph, which is a necessary property for exact recovery in the stochastic block model (SBM) [31]. Furthermore, motivated by the social homophily theory [25], we assume $\alpha \geq \beta \geq \gamma$ where users within the same group (or cluster) are more likely to be connected than those in different groups (or clusters).

Let ψ denote an estimator (decoder) that takes as input a pair $(Y, \mathcal{G})$, and outputs a completed rating matrix $X \in \mathbb{F}_q^{n \times m}$. Note that both the set of rating vectors $\mathcal{V}$ and the user partitioning $\mathcal{Z}$ can be recovered from the completed rating matrix X and vice versa. With a slight abuse of notation, we denote the output of the estimator as X or $(\mathcal{V}, \mathcal{Z})$.

A key parameter of the main result (see Section III) is the discrepancy between the rating vectors. Let δ_g be the minimum normalized Hamming distance among the distinct pairs of rating vectors of groups *within the same cluster*. Let δ_c be the counterpart with respect to different pairs of rating vectors *across different clusters*. Formally, δ_g and δ_c are given by

$$\delta_g = \frac{1}{m} \min_{x \in [c]} \min_{\substack{i,j \in [g] \\ i \neq j}} d_{\mathrm{H}}\left(u_i^{(x)}, u_j^{(x)}\right), \quad \delta_c = \frac{1}{m} \min_{\substack{i,j \in [g] \\ x,y \in [c], x \neq y}} d_{\mathrm{H}}\left(u_i^{(x)}, u_j^{(y)}\right).$$

Our main result hinges on $\delta := (\delta_g, \delta_c)$. We provide theoretical guarantees for the recovery of all rating matrices M in which the rating vectors maintain a minimum level of dissimilarity. Formally, define $\mathcal{M}^{(\delta)}$ as the set of matrices $M = (\mathcal{V}, \mathcal{Z})$ such that the following properties are satisfied: (i) the set of rating vectors $\mathcal{V}$ must satisfy the property that the minimum normalized Hamming distance among rating vectors in different groups within the same cluster and those in different clusters are not smaller than δ_g and δ_c, respectively; and (ii) the user partitioning $\mathcal{Z}$ must satisfy the property that the sizes of clusters and groups are c/n and $c/(ng)$ users, respectively.

The performance metric we consider to provide theoretical guarantees on the recommendation quality is the worst-case probability of error P_e. In other words, the quality of the estimator is defined by its accuracy of estimating the *most difficult* ground truth matrix $M = (\mathcal{V}, \mathcal{Z}) \in \mathcal{M}^{(\delta)}$. Therefore, we apply a minimax optimization approach wherein the objective is to find the estimator that minimizes the maximum risk, i.e.,

$$\inf_{\psi} P_e^{(\delta)}(\psi) = \inf_{\psi} \max_{M \in \mathcal{M}^{(\delta)}} \mathbb{P}\left[\psi(Y, \mathcal{G}) \neq M\right]. \tag{1}$$

Our goal is to characterize the optimal sample complexity (i.e., the minimum number of entries of the rating matrix that is required to be observed), concentrated around $nmp^\star$ in the limit of n and m, for exact rating matrix recovery. Here, $p^\star$ denotes a sharp threshold on the observation probability such that the following conditions, in the limit of n and m, are satisfied: (i) when $p > p^\star$, there exists an estimator such that the error probability can be made arbitrarily close to 0; and (ii) when $p < p^\star$, the error probability does not converge to zero no matter what and whatsoever.

III. Main Results

Similar to [21], [32], we assume that $m = \omega(\log n)$ and $\log m = o(n)$ in order to apply large deviation theories. These assumptions are practically relevant, as they eliminate the possibility of having extremely tall or wide matrices.

Theorem 1 (Optimal Sample Complexity)**.** *Let $m = \omega(\log n)$ and $\log m = o(n)$. Let q, θ, c, g and r be constants such that q is prime, $\theta \in [0, (q-1)/q)$, and $r \leq g$. Let $\Upsilon(q,\theta) := 1/(\sqrt{1-\theta} - \sqrt{\theta/(q-1)})^2$. For any constant $\epsilon > 0$, if* (2) *holds, then there exists an estimator ψ that outputs a rating matrix $X \in \mathcal{M}^{(\delta)}$ given Y and $\mathcal{G}$ such that $\lim_{n\to\infty} P_e^{(\delta)}(\psi) = 0$; conversely, if* (3) *holds, then $\lim_{n\to\infty} P_e^{(\delta)}(\psi) \neq 0$ for any estimator ψ. Therefore, the optimal observation probability $p^\star$ is given by* (4)*.*

Proof: We provide a proof sketch of Theorem 1. We defer the complete achievability and converse proofs to [33]. The achievability proof is based on maximum likelihood estimation (MLE). We first evaluate the likelihood for a given clustering/grouping of users and the corresponding rating matrix. Next, we provide an upper bound on the worst-case probability of error, which is given by the probability that the likelihood of the ground truth rating matrix is less than that of a candidate rating matrix. Then, we partition the candidate rating matrices into two sets, typical and atypical sets. A typical (or atypical) set denotes the set of rating matrices that have a relatively small (or large) number of error entries compared to the ground truth matrix. Finally, we conduct typical and atypical error analyses as follows. In the typical error analysis, we provide a tight upper bound on the cardinality of the typical set and a loose upper bound on the error probability of a candidate matrix. On the other hand, in the atypical error analysis, we provide a loose upper bound on the cardinality of the typical set and a tight upper bound on the error probability of a candidate matrix. These analyses are based on the fact that the size of the set of candidate matrices with a small number of error entries is relatively larger than that of the one with a large number of error entries. Based on these bounds, we show that the probability of error for any candidate matrix in the typical set is negligibly smaller than the carnality of the typical set of matrices, and hence this leads to convergence of the overall worst-case probability of error to zero as n and m goes to infinity. Hence, the worst-case probability of error vanishes in the limit of n and m. This completes the achievability proof.

The converse proof starts with establishing a lower bound on the error probability and showing that it is minimized when employing the maximum likelihood estimator. Next, we prove that if p is smaller than any of the three terms in the RHS of (4), then there exists another solution that yields a larger likelihood, compared to the ground-truth matrix. More precisely, for any estimator and any ground truth rating matrix, we present the following three cases. First, if $p \leq \Upsilon(q,\theta)\frac{(1-\epsilon)gc\log m}{(g-r+1)n}$, there exists a class of matrices obtained by replacing one column of the ground truth rating matrix with a carefully chosen sequence, and it yields a higher likelihood than that of the ground truth rating matrix. Second, if $p \leq \Upsilon(q,\theta)\frac{\log n}{\delta_g m}((1-\epsilon) - \frac{(\sqrt{\alpha}-\sqrt{\gamma})^2}{gc})$, there exists a class of rating matrices obtained by swapping the rating vectors of two users in the same cluster yet from distinct groups with the Hamming distance between their rating vectors being $m\delta_g$. We show that the likelihood of any rating matrix from this class is greater than the one of the ground truth rating matrix. Third, if $p \leq \Upsilon(q,\theta)\frac{\log n}{\delta_c m}((1-\epsilon) - \frac{(\sqrt{\alpha}-\sqrt{\gamma})^2+(g-1)(\sqrt{\beta}-\sqrt{\gamma})^2}{gc})$, we can find a class of rating matrices obtained by swapping the rating vectors of two users in distinct clusters with an $m\delta_c$ Hamming distance between their rating vectors. We show that any rating matrix from this class yields a larger likelihood than that of the ground truth rating matrix. For each case, we show that the maximum likelihood estimator will fail in the limit of n and m by selecting one of the rating matrices from the respective class instead of the ground truth rating matrix. This completes the converse proof and concludes the proof of Theorem 1. ■

Remark 1. *The technical distinctions from the previous works [21], [22], [32] are four-fold. First, the likelihood computation requires more involved combinatorial arguments due to the hierarchical structure of the similarity graph (See Lemma 1 in [33]). Second, sophisticated upper and lower bounding techniques are developed to leverage the relational structure across different groups (See Lemmas 3 and 4 in [33]). Third, novel typical and atypical error analyses are proposed for the achievability proof (See Lemmas 5 and 6 in [33]). Fourth, novel failure proof techniques are developed for the converse proof (See Section V in [33]). Furthermore, setting $(c,g,r,q)=(2,3,2,2)$, the optimal probability $p^\star$ in* (4) *reduces to the one characterized by [32]. Hence, the result of [32] is subsumed by the general result given by Theorem 1. Note that the optimal sample complexity for general (c,g,r,q) is conjectured by [32]. However, the proofs are provided only for $(c,g,r,q)=(2,3,2,2)$. This work presents complete achievability and converse proofs for any (c,g,r,q).*

$$p \geq \Upsilon(q,\theta)\ \max\left\{\frac{gc(1+\epsilon)}{g-r+1}\frac{\log m}{n}, \frac{\log n}{\delta_g m}\left((1+\epsilon) - \frac{\left(\sqrt{\alpha}-\sqrt{\beta}\right)^2}{gc}\right), \frac{\log n}{\delta_c m}\left((1+\epsilon) - \frac{\left(\sqrt{\alpha}-\sqrt{\gamma}\right)^2+(g-1)\left(\sqrt{\beta}-\sqrt{\gamma}\right)^2}{gc}\right)\right\}. \quad (2)$$

$$p \leq \Upsilon(q,\theta)\ \max\left\{\frac{gc(1-\epsilon)}{g-r+1}\frac{\log m}{n}, \frac{\log n}{\delta_g m}\left((1-\epsilon) - \frac{\left(\sqrt{\alpha}-\sqrt{\beta}\right)^2}{gc}\right), \frac{\log n}{\delta_c m}\left((1-\epsilon) - \frac{\left(\sqrt{\alpha}-\sqrt{\gamma}\right)^2+(g-1)\left(\sqrt{\beta}-\sqrt{\gamma}\right)^2}{gc}\right)\right\}. \quad (3)$$

$$p^\star = \Upsilon(q,\theta)\ \max\left\{\frac{gc}{g-r+1}\frac{\log m}{n}, \frac{\log n}{\delta_g m}\left(1 - \frac{\left(\sqrt{\alpha}-\sqrt{\beta}\right)^2}{gc}\right), \frac{\log n}{\delta_c m}\left(1 - \frac{\left(\sqrt{\alpha}-\sqrt{\gamma}\right)^2+(g-1)\left(\sqrt{\beta}-\sqrt{\gamma}\right)^2}{gc}\right)\right\}. \quad (4)$$

IV. Discussion

A. Noise Model and Finite Field Size

The reason for choosing the uniform noise model is that the uniform noise distribution is the worst case distribution in discrete channels. Next, it is evident that the optimal sample complexity in (4) increases as θ increases. Furthermore, as θ approaches $(q-1)/q$, each sampled entry of the rating matrix can take any of the q possible values with a uniform probability of $1/q$, and hence an infinite sample complexity is theoretically required to exactly recover the entries of the rating matrix.

B. Quality of Hierarchical Similarity Graph

We illustrate the relationship between the optimal sample complexity and the quality of the hierarchical graph by defining the following quality parameters: $I_{\alpha,\beta} \coloneqq \left(\sqrt{\alpha}-\sqrt{\beta}\right)^2$, $I_{\alpha,\gamma} \coloneqq \left(\sqrt{\alpha}-\sqrt{\gamma}\right)^2$, and $I_{\beta,\gamma} \coloneqq \left(\sqrt{\beta}-\sqrt{\gamma}\right)^2$. Intuitively, as $I_{\alpha,\beta}$ increases, it becomes easier to distinguish users in different groups within the same cluster. On the other hand, higher values of $I_{\alpha,\gamma}$ and $I_{\beta,\gamma}$ lead to better user clustering. The optimal sample complexity reads different values depending on the quality parameters of the hierarchical graph. Next, we define three regimes as follows. The first term in the RHS of (4) is activated when $I_{\alpha,\beta}, I_{\alpha,\gamma}$ and $I_{\beta,\gamma}$ are large enough so that the grouping and clustering information is reliable. Therefore, this regime is coined as *"perfect clustering/grouping regime"*. The second term in the RHS of (4) is activated when $I_{\alpha,\beta}$ is small such the grouping information is not reliable, Therefore, this regime is coined as *"grouping-limited regime"*. The third term in the RHS of (4) is activated when $I_{\alpha,\gamma}$ and $I_{\beta,\gamma}$ are small such that the clustering information is not reliable, and $\delta_g > \delta_c$. Thus, this regime is coined as *"clustering-limited regime"*. Next, we analyze the optimal sample complexity under each regime. For illustrative purposes, we assume that $\theta = 0$. This implies $\Upsilon(q,\theta) = 1$.

1) Perfect Clustering/Grouping Regime: The optimal sample complexity reads $(gc/(g-r+1))m\log m$. Since the grouping and clustering information are reliable, one can recover the groups and clusters from the similarity graph. However, further increments of the values of these quality parameters do not yield further improvement in the sample complexity, and hence the sample complexity gain from the similarity graph is saturated in this regime. Moreover, it should be noted that a naive generalization of [21], [22] requires $crm\log m$ observations since there are r independent rating vectors to be estimated for each the c clusters, and each rating vector requires $m\log m$ observations under the considered random sampling due to the coupon-collecting effect. On the other hand, we leverage the relational structure (i.e, linear dependency) across the rating vectors of different groups, reflected by the underlying linear MDS code structure (See Section IV in [33]), and hence this serves to estimate the rc rating vectors more efficiently, precisely by a factor of $r(g-r+1)/g$ improvement, thus yielding $(gc/(g-r+1))m\log m$.

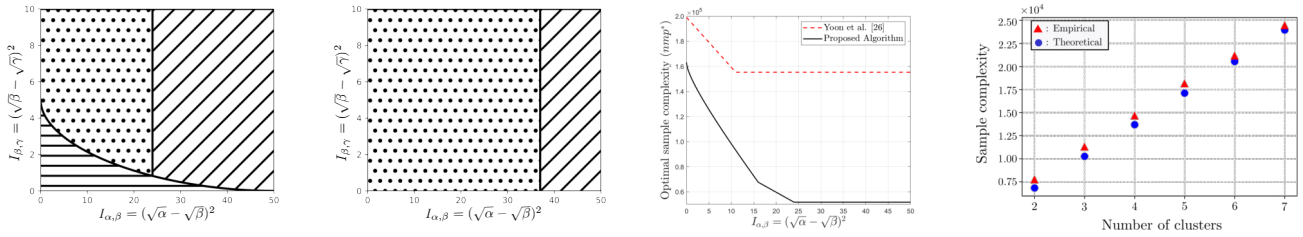
2) Grouping-Limited Regime: The optimal sample complexity reads $\frac{1}{\delta_g}\left(1-\frac{I_{\alpha,\beta}}{gc}\right)n\log n$, which is a decreasing function of $I_{\alpha,\beta}$. This sample complexity coincides with that of [22] in which the considered similarity graph consists of only gc clusters. This implies that leveraging the relational structure across different groups does not help improve the sample complexity when the grouping information is not reliable. Moreover, since the clustering information is reliable, the clusters can be recovered from the similarity graph. However, further increases in $I_{\alpha,\gamma}$ and $I_{\beta,\gamma}$ do not reduce sample complexity, so the sample complexity gain from these two parameters is saturated in this regime.

3) Clustering-Limited Regime: The optimal sample complexity reads $\frac{1}{\delta_c}\left(1-\frac{I_{\alpha,\gamma}+(g-1)I_{\beta,\gamma}}{gc}\right)n\log n$, which is a decreasing function of $I_{\alpha,\gamma}$ and $I_{\beta,\gamma}$. This is the most challenging scenario, which has not been explored by any prior works. Since the clustering information is not reliable, it is not possible to recover the groups and clusters from the similarity graph. Furthermore, note that when $\beta = \gamma$, i.e., groups and clusters are indistinguishable, we have $I_{\alpha,\beta} = I_{\alpha,\gamma}$ and $I_{\beta,\gamma} = 0$. As a result, it boils down to a problem setting of gc clusters, and hence the optimal sample complexity reads $\frac{1}{\delta_c}\left(1-\frac{I_{\alpha,\beta}}{gc}\right)n\log n$. Comparing to the optimal sample complexity expression for the grouping-limited regime, the only distinction appears in the denominator, in which δ_g is replaced with δ_c due to the fact that $\delta_c < \delta_g$.

C. Illustrative Comparisons

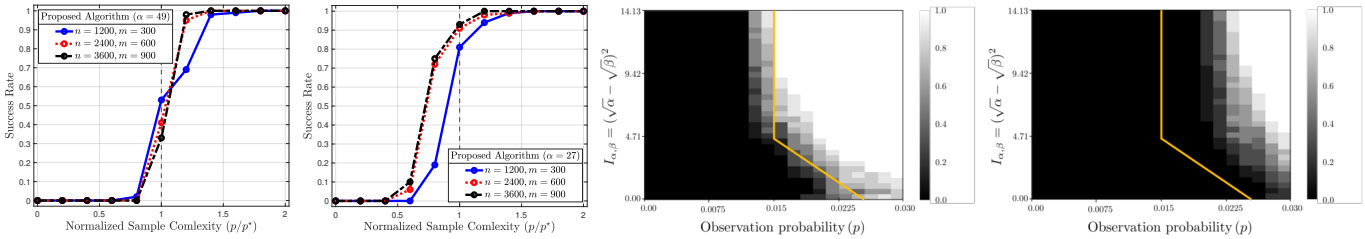
Fig. 1a and Fig. 1b depict the different regimes of the optimal sample complexity as a function of $(I_{\alpha,\beta}, I_{\beta,\gamma})$. In Fig. 1a, where $\delta_g > \delta_c$, the region depicted by diagonal stripes corresponds to the perfect clustering/grouping regime and the first term in the RHS of (4) is active. The graph quality parameters $I_{\alpha,\beta}$, $I_{\beta,\gamma}$, and consequently $I_{\alpha,\gamma}$ are large, and the graph information is rich enough to perfectly retrieve the clusters and groups. The region represented by dots corresponds to the grouping-limited regime, where the second term in the RHS of (4) is active. In this regime, graph information suffices to exactly recover the clusters, but we need to rely on rating observation to exactly recover the groups. Finally, the third term in the RHS of (4) is active in the region captured by horizontal stripes. This indicates the clustering-limited regime, where neither clustering nor grouping is exact without the side information of the rating vectors. On the other hand, Fig. 1b, where $\delta_g < \delta_c$, depicts the practically-relevant setting in which the rating vectors of users in the same cluster are expected to be more similar than those in different clusters. Note that the third regime (clustering-limited regime) vanishes in Fig. 1b.

Fig. 1c compares the optimal sample complexity, as a function of $I_{\alpha,\beta}$, between the one reported by Theorem 1 and that of [22]. Note that [22] leverages neither the hierarchical structure of the graph nor the linear dependency among the rating vectors. Thus, the problem formulated in Section II will be translated to a graph that consists of gc clusters whose rating vectors are linearly independent in the setting



(a) $\delta_g = 1/3$ and $\delta_c = 1/6$. (b) $\delta_g = 1/6$ and $\delta_c = 1/3$. (c) $\delta_g = 1/3$ and $\delta_c = 1/6$. (d) $\delta_g = \delta_c = 1/3$.

Figure 1. Let $(n, m, \theta, c, g, r, q) = (4000, 500, 0, 10, 5, 3, 5)$ for subfigures (a)-(c). (a), (b): The different regimes of the optimal sample complexity reported in Theorem 1 for $\delta_g > \delta_c$ and $\delta_g < \delta_c$, respectively. Diagonal stripes, dots, and horizontal stripes refer to perfect clustering/grouping regime, grouping-limited regime, and clustering-limited regime, respectively. (c) Comparison between the sample complexity reported in Theorem 1 and that of [22] for $\beta = 5$ and $\gamma = 1$. (d) Sample complexity as a function of the number of clusters in the perfect clustering/grouping regime where $(n, m, \theta, g, r, q) = (1680, 840, 4, 0, 3, 5)$ and $(\alpha, \beta, \gamma) = (49, 9, 0.5)$.



(a) Perfect Clustering/Grouping Regime (b) Grouping-Limited Regime (c) Proposed Algorithm (d) Algorithm of [22]

Figure 2. (a), (b): The success rate of the proposed algorithm as a function of $p/p^\star$ for different values of n, m and α. The problem setting is $(c, g, q, r) = (3, 4, 5, 3)$, $\theta = 0.01$, $(\beta, \gamma) = (9, 0.5)$, and $(\delta_g, \delta_c) = (1/3, 1/3)$. The MDS code structure is $u_4^{(x)} = u_1^{(x)} + u_2^{(x)} + u_3^{(x)}$ for $x \in [3]$. (c), (d): The success rates of two matrix completion algorithms where $(n, m, \theta, \gamma, c, g, r, q) = (2400, 600, 0, 0.5, 3, 4, 3, 5)$, and $(\delta_g, \delta_c) = (1/2, 1/2)$.

of [22]. Furthermore, the minimum Hamming distance for [22] is δ_c. The significant gain in the sample complexity of our result is evident in the diagonal parts of the plot (i.e., clustering-limited and grouping-limited regimes on the left side) is due to leveraging the hierarchical graph structure, while the improvement in the sample complexity in the flat part of the plot (i.e, perfect clustering/grouping regime) is a consequence of leveraging the relational structure (i.e., linear dependency) among the rating vectors within each cluster.

Fig. 1d depicts the sample complexity as a function of the number of clusters in the perfect clustering/grouping regime. It shows both the theoretical values (given by (4)) and the empirical values (given by the algorithm explained in Section V). It is evident that the sample complexity increases linearly with the number of clusters when there is enough graph side information to retrieve the cluster and group structures.

V. Simulation Results

We conduct Monte Carlo experiments to show that the proposed algorithm achieves p^* characterized by Theorem 1. Empirical success rates are averaged on 100 random realizations of rating vectors and hierarchical graphs. The settings of the experiments are stated in the captions of the figures. The proposed algorithm is built in part upon the computationally efficient matrix completion algorithm proposed in [32]. It consists of four phases. The first one exactly recovers the clusters using the community detection algorithm in [34]. The second phase gives an initial estimate of the groups (i.e, almost exact recovery) using any spectral clustering algorithm, e.g. [31], [35]–[39]. Next, the third phase exactly recovers the rating vectors associated with each group in each cluster using maximum likelihood estimation. Finally, the last phase exactly recovers the group via an iterative local refinement procedure. The distinction of our algorithm compared to [32] is that the stage of exact recovery of the rating vectors is based on maximum likelihood estimation.

In Figs. 2a and 2b, we quantify the empirical success rate of the proposed algorithm as a function of the normalized sample complexity. We vary n and m such that $n/m = 4$. Fig. 2a shows the case of α=49 which corresponds to perfect clustering/grouping regime, while Fig. 2b depicts the case of α=27 which corresponds to the grouping-limited regime. In both figures, we observe a phase transition in the success rate at $p = p^\star$, and the phase transition gets sharper as n and m increase. This implies that the proposed algorithm achieves p^*, given by Theorem 1, in different regimes when the graph side information is not scarce.

Finally, we highlight the sample complexity gain from leveraging the relational structure among rating vectors. Figs. 2c and 2d depict the success rates under various values of p and $I_{\alpha,\beta}$ for the proposed algorithm and the one in [22] (where the relational structure among rating vectors is not considered), respectively. The empirical success rate is represented by a grayscale heat map. The orange line indicates the optimal sample complexity given by Theorem 1. The vertical and diagonal lines correspond to the sample complexity in perfect clustering/grouping and grouping-limited regimes, respectively. In Fig. 2c, the phase transition in the success rate of the proposed algorithm is sharp and occurs at the optimal probability given by (4). However, the phase transition in Fig. 2d occurs at a higher observation probability, and therefore [22] requires a higher sample complexity than the proposed algorithm.

REFERENCES

[1] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

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

[3] 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.

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

[5] W.-J. Li and D.-Y. Yeung, "Relation regularized matrix factorization," *Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

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

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

[8] H. Ma, H. Yang, M. R. Lyu, and I. King, "SoRec: social recommendation using probabilistic matrix factorization," *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 931–940, 2008.

[9] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 203–210, 2009.

[10] 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," *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] H. Zhao, Q. Yao, J. T. Kwok, and D. L. Lee, "Collaborative filtering with social local models," *IEEE International Conference on Data Mining (ICDM)*, pp. 645–654, 2017.

[12] S. Chouvardas, M. A. Abdullah, L. Claude, and M. Draief, "Robust online matrix completion on graphs," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4019–4023, 2017.

[13] P. Massa and P. Avesani, "Controversial users demand local trust metrics: An experimental study on epinions. com community," *AAAI*, pp. 121–126, 2005.

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

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

[16] ——, "Using a trust network to improve top-n recommendation," *Proceedings of the third ACM conference on Recommender systems*, pp. 181–188, 2009.

[17] 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.

[18] X. Yang, H. Steck, Y. Guo, and Y. Liu, "On top-k recommendation using social networks," *Proceedings of the sixth ACM conference on Recommender systems*, pp. 67–74, 2012.

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

[20] R. v. d. Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," *arXiv preprint arXiv:1706.02263*, 2017.

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

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

[23] Q. Zhang, V. Y. F. Tan, and C. Suh, "Community detection and matrix completion with social and item similarity graphs," *IEEE Transactions on Signal Processing*, vol. 69, pp. 917–931, 2021.

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

[25] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[26] 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.

[27] V. Y. Tan, L. Balzano, and S. C. Draper, "Rank minimization over finite fields: Fundamental limits and coding-theoretic interpretations," *IEEE transactions on information theory*, vol. 58, no. 4, pp. 2018–2039, 2011.

[28] A. Emad and O. Milenkovic, "Information theoretic bounds for tensor rank minimization over finite fields," *IEEE Global Telecommunications Conference (GLOBECOM)*, 2011.

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

[30] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe, "Community detection and classification in hierarchical stochastic blockmodels," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 1, pp. 13–26, 2016.

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

[32] A. Elmahdy, J. Ahn, C. Suh, and S. Mohajer, "Matrix completion with hierarchical graph side information," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[33] J. Ahn, A. Elmahdy, S. Mohajer, and C. Suh, "On the fundamental limits of matrix completion: Leveraging hierarchical similarity graphs," *arXiv preprint arXiv:2109.05408*, 2021.

[34] 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.

[35] 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 (JMLR)*, vol. 18, no. 1, pp. 1980–2024, 2017.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[37] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems (NIPS)*, pp. 849–856, 2002.

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

[39] J. Lei, A. Rinaldo *et al.*, "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, vol. 43, no. 1, pp. 215–237, 2015.