Collaborative Clustering: Sample Complexity and Efficient Algorithms

Jungseul Ok[*] JUNGSEUL@KTH.SE
Se-Young Yun[†] YUNSEYOUNG@KAIST.AC.KR
Alexandre Proutiere[*] ALEPRO@KTH.SE
Rami Mochaourab[*] RAMIMO@KTH.SE
[*] *KTH, The Royal Institute of Technology, Stockholm, Sweden*
[†] *Korea Advanced Institute of Science and Technology, Daejeon, South Korea*

Editors: Steve Hanneke and Lev Reyzin

Abstract

We study the problem of *collaborative clustering*. This problem is concerned with a set of items grouped into clusters that we wish to recover from ratings provided by users. The latter are also clustered, and each user rates a random but typical small number of items. The observed ratings are random variables whose distributions depend on the item and user clusters only. Unlike for collaborative filtering problems where one needs to recover both user and item clusters, here we only wish to classify items. The number of items rated by a user can be so small that anyway, estimating user clusters may be hopeless. For the collaborative clustering problem, we derive fundamental performance limits satisfied by any algorithm. Specifically, we identify the number of ratings needed to guarantee the existence of an algorithm recovering the clusters with a prescribed level of accuracy. We also propose SplitSpec, an algorithm whose performance matches these fundamental performance limit order-wise. In turn, SplitSpec is able to exploit, as much as this is possible, the users' structure to improve the item cluster estimates.

Keywords: Collaborative clustering, sample complexity, spectral method

1. Introduction

Cluster analysis consists of dividing a set of items into a small number of meaningful and useful groups based on the data that describe the items. In its classical form, the item description comes in the form of a feature vector, and with such data, clustering can be efficiently performed for instance using the celebrated k-means algorithm or its variants (if properly initialized) (Bachem et al., 2016). In this paper, we investigate a clustering task where items are described by labels or ratings independently provided by users. We assume here that both items and users are clustered, in the sense that the rating statistics of an item by a given user only depend on the item and user clusters. Our clustering task has similarities with Collaborating Filtering (CF) (Ekstrand et al., 2011), a critical tool used in recommender systems. However, CF aims at predicting the unobserved ratings, i.e., at assessing whether a user would like or dislike an item. CF hence often reduces to a matrix completion problem, which in turn requires to be able to estimate the clusters of both items and users.

In contrast here, we only wish to recover the item clusters. In particular, we may consider scenarios where users provide ratings only a very small subset of items, making it almost impossible to efficiently reconstruct user clusters. Nevertheless, our objective is to exploit as much as we can the users' structure to accurately estimate the item clusters. We refer to this task as *Collaborative Clustering*. This term is not new, and has been used in other contexts (Yue et al., 2014); but we believe that it faithfully captures the nature of our problem.

For the collaborative clustering problem, we derive fundamental sample complexity lower bounds, expressing the number of ratings that any algorithm (even optimal) would require to cluster items with a given prescribed error rate. We also present SplitSpec, a clustering algorithm that achieves this fundamental performance limit order-wise. SplitSpec hence efficiently exploits users' structure, and performs well in regimes where typical CF algorithms would fail (if the number of ratings per user is low) and where classical clustering algorithms such k-means and its variants can hardly discover the items' structure (if the ratings of an item averaged over users do not depend on the cluster of this item).

1.1. Model and Objectives

Consider a set of n items $\mathcal{V} = [n] := \{1, \dots, n\}$ partitioned into a set of K disjoint clusters $\mathcal{V}_1, \dots, \mathcal{V}_K$, i.e., $\bigcup_{i=1}^{K} \mathcal{V}_i = \mathcal{V}$ and $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$, for all $i \neq j$. Each item $v \in \mathcal{V}$ is assigned to a cluster $\mathcal{V}_i$ with probability $\alpha_i > 0$ independently of other items. The number K of clusters and the distribution $\alpha = (\alpha_1, ..., \alpha_K)$ of items into clusters are assumed not to depend on the number of items n, so that each cluster has a size linearly growing with n in average. Without loss of generality, we further assume that $\alpha_1 \leq \alpha_2 \cdots \leq \alpha_K$. The objective is to recover the clusters by collecting and analyzing the "ratings" of items provided by a set $\mathcal{U} = [m]$ of users. The latter can be categorized into L types, where each type defines the rating statistics for items in the various clusters. Any given user is of type ℓ with probability $\beta_\ell > 0$, chosen independently of other users. The distribution $\beta = (\beta_1, ..., \beta_L)$ does not depend on the number of items n, nor on the number of users m. The types of the various users are assumed to be unknown.

Rating statistics. For any item-cluster i and user-type ℓ, if we have access to the rating X_{uv} of item $v \in \mathcal{V}_i$ by a type-ℓ user, then $X_{uv} = 1$ (like) with probability $p_{\ell i}$, and $X_{uv} = 0$ (dislike) with probability $1 - p_{\ell i}$. The parameter $p = (p_{\ell i})_{\ell \in [L], i \in [K]}$ defining the rating statistics is unknown, and does not depend on n nor m. We make the following mild assumption: (A1) there exists a constant $\eta \geq 1$ such that for every $i, j \in [K]$ and $\ell \in [L]$, $\frac{p_{\ell i}}{p_{\ell j}} \leq \eta$ and $\frac{1-p_{\ell i}}{1-p_{\ell j}} \leq \eta$. (A1) just states some homogeneity in the average ratings of items across clusters.

Available ratings. Each user provides ratings for at most w items. We consider two scenarios:
(i) Random assignment: the w items assigned to a given user are chosen uniformly at random, and independently of the items assigned to the other users. Each user $u \in [m]$ is first assigned a set $\mathcal{W}_u$ of w items chosen uniformly at random. Then user u rates each item in $\mathcal{W}_u$ with probability $q > 0$, independently of other items. The set of items for which user u provided ratings is denoted by $\mathcal{R}_u$. The latter is of average cardinality wq. w and q may depend on n and m. This model is simple, and yet captures some randomness in the

number of items rated by the various users. We will often make the following assumption: (A2) $wq \geq 1$. (A2) holds as soon as each user provides at least one rating, which seems very reasonable (we can remove from the dataset users who do not provide any rating). We will however also discuss the case where $wq < 1$.
(ii) Adaptive assignment: users are here supposed to rate items sequentially, and we have the ability to choose the w items to be rated by the current user depending on the ratings provided by the previous users. However, the paper mainly focuses on the random assignment scenario.

In summary, the model on the way ratings are generated and made available is parameterized by α, β, p, w and q. To simplify the notations, we denote by $\mathcal{M} = (\alpha, \beta, p, w, q)$ the model parameters.

Clustering algorithms and their performance. In the random assignment scenario, a clustering algorithm π takes as input the available ratings and output a set of disjoint clusters $\hat{\mathcal{V}}_1, \dots, \hat{\mathcal{V}}_K$. In the adaptive assignment scenario, the algorithm also decides on the items to be rated by each user (depending on ratings provided by previous users). In both cases, the performance of the algorithm π is assessed through the number $\varepsilon^\pi(n,m)$ of items that it misclassifies:

$$\varepsilon^\pi(n,m) = \min_\theta \sum_{k=1}^K \left| \hat{\mathcal{V}}_k \setminus \mathcal{V}_{\theta(k)} \right|$$

where the minimum is taken over all permutations θ of $[K]$, and where for any set A, $|A|$ denotes the cardinality of A.

1.2. Main Results

In this paper, we derive fundamental performance limits satisfied by any clustering algorithm under random and adaptive assignments (Theorems 1 and 2, respectively). More precisely, we provide a lower bound on the *sample complexity* of our clustering problem. The sample complexity is defined as the number of ratings (cumulated over users) required to get an accurate cluster detection. We also present SplitSpec, a clustering algorithm that achieves these limits order-wise (Theorem 3).

Sample complexity – Random assignment. To formalize our misclassification lower bound, we need to introduce the notion of *inter-cluster divergence* $D_{ij}(\mathcal{M})$ for clusters $i \neq j$. This divergence, precisely defined in Section 3.1, depends on the model parameters $\mathcal{M} = (\alpha, \beta, p, w, q)$ and characterizes the hardness of distinguishing items from two clusters i and j solely based on the available ratings. Small value of $D_{ij}(\mathcal{M})$ indicates that items from clusters i and j are difficult to distinguish. The next theorem establishes the direct connection between the average number of misclassified items and the minimal inter-cluster divergence defined as $D(\mathcal{M}) = \min_{i \neq j} D_{ij}(\mathcal{M})$.

Theorem 1 *Assume that $\mathcal{M}$ satisfies Assumption (A1), and consider the random assignment scenario. Further assume that there exists a clustering algorithm π that misclassifies at most $s = o(n)$ items on average, i.e., $\limsup_{n\to\infty} \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{s} \leq 1$. Then, when $wq = o(\sqrt{n})$,*

$$\liminf_{n\to\infty} \frac{mwqD(\mathcal{M})}{n \log(n/s)} \geq 1 \ .$$

Furthermore, when $wq = \Omega(\sqrt{n})$, $mwq = \Omega(n \log n/s)$.

We make the following important remarks on the above theorem.
1. The theorem implies that to be able to misclassify s items only, one needs to get the ratings from at least $n\log(n/s)/(wqD(\mathcal{M}))$ users, and hence a sample complexity greater than $n\log(n/s)/D(\mathcal{M})$ (the average number ratings each user is providing is wq).
2. Another important consequence of the theorem is that the number of misclassified items will grow linearly with the number of items, irrespective of the number of available ratings, if and only if $D(\mathcal{M}) = 0$ – in which case we say that there are *indistinguishable* clusters. The condition for indistinguishability is simple and depends on whether the maximum number of ratings per user w is equal to or larger than 1 (see Proposition 4): when $w = 1$, $D(\mathcal{M}) = 0$ iff $\exists i \neq j : \sum_\ell \beta_\ell (p_{\ell i} - p_{\ell j}) = 0$, and when $w \geq 2$, $D(\mathcal{M}) = 0$ iff $\exists i \neq j : \forall \ell, p_{\ell i} = p_{\ell j}$. In other words, when the number of ratings per user is at most 1, then two clusters become indistinguishable as soon as the average ratings of their items are the same. However when $w \geq 2$, for two clusters to be indistinguishable, their items need to have exactly the same rating statistics across all user types. Theorem 1 then implies that as soon as $w \geq 2$, we can separate clusters unless they have the same rating statistics.
3. When $wq \geq 1$ and $D(\mathcal{M}) > 0$, i.e., when item clusters are distinguishable, it is easy to check that $D(\mathcal{M})$ is upper and lower bounded by a constant that does not depend on the model parameters $\mathcal{M}$ (see Proposition 5). Hence, our lower bound on the sample complexity is $Cn\log(n/s)$, where C is a constant that does not depend on n, m, and the model parameters $\mathcal{M}$.
4. As a final remark, observe that when $D(\mathcal{M}) > 0$, the sample complexity exhibits the same scaling in n, m, and s irrespective of the users' structure. This scaling does not depend on the number of user types. This comes as a surprise because it means that as long as $w \geq 2$, we cannot hope to improve the sample complexity (order-wise) by leveraging users' structure. We will make this statement more precise in Section 3.

Sample complexity – Adaptive assignment. We derive a similar lower bound for scenarios where the items to be rated can be sequentially selected. To this aim, we need to introduce the notion of inter-cluster maximal divergence $\tilde{D}_{ij}(\mathcal{M})$, see Section 3.3 for a formal definition. We also define $\tilde{D}(\mathcal{M}) = \min_{i\neq j} \tilde{D}_{ij}(\mathcal{M})$.

Theorem 2 *Assume that* $\mathcal{M}$ *satisfies Assumption (A1), and consider the adaptive assignment scenario. Further assume that there exists a sequential item selection and clustering algorithm* π *that misclassifies at most* $s = o(n)$ *items on average, i.e.,* $\limsup_{n\to\infty} \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{s} \leq 1$. *Then, when* $wq = o(\sqrt{n})$,

$$\liminf_{n\to\infty} \frac{mwq\tilde{D}(\mathcal{M})}{(\alpha_1/3)n\log(n/s)} \geq 1 \ .$$

We will see that $\tilde{D}_{ij}(\mathcal{M})/\alpha_1 \geq \tilde{D}_{ij}(\mathcal{M}) \geq D_{ij}(\mathcal{M})$, and hence being able to select the items that each user rates in an adaptive manner naturally improves our sample complexity lower bound. Note however that surprisingly, this improvement can be by a constant factor only, where the lower bound retains the same scaling in n, w, and q as the lower bound in the random assignment scenario.

The SplitSpec algorithm and its efficiency. To estimate the item clusters from the available ratings, we propose SplitSpec, a clustering algorithm that proceeds in two steps: (i) in the Split step, it constructs from the available ratings an undirected random weighted graph whose nodes are items. The ratings of each user u are used to add edges in the graph. To do so, we randomly split the set $\mathcal{W}_u$ of user-u's potential ratings into subgroups. Each subgroup can then generate an edge in the graph. (ii) In its second step, the algorithm applies a spectral decomposition of the weight matrix of the graph constructed in the Split step, and uses the decomposition to estimate the item clusters.

The critical step of SplitSpec is the way the graph is constructed. Our construction strikes a good trade-off between two objectives: the graph should capture the information of the available ratings as faithfully as possible; and the entries of the resulting weight matrix should be as stochastically independent as possible to allow us to analyze the performance of the algorithm (leveraging results from random matrix theory). Ideally if all the ratings of a given user were jointly used to construct the graph, we would need to consider graphs with *hyper-edges*, where an hyper-edge would be added between all items positively rated by the user. However with hyperedges, the weight matrix entries would not be independent, which would make the spectral analysis cumbersome.

The following theorem provides a performance guarantee for SplitSpec when (A2) holds (i.e., when $wq \geq 1$). In Section 4, we also analyze its performance when $wq < 1$.

Theorem 3 *Consider the collaborative model $\mathcal{M}$ with the random assignment of items. Assume that $w \geq 2$, $wq = o(\sqrt{n})$, $(wq)^2 m = o(n^2)$, $D(\mathcal{M}) > 0$, and that* (A1) *and* (A2) *hold. For any given $s = o(n)$, there exists a constant $C > 0$ such that under the* SplitSpec *algorithm, when $mwq \geq Cn \log(n/s)$, the number of misclassified items is less than s with high probability.*

In view of the above theorem, SplitSpec exhibits an order-optimal sample complexity for any $s = o(n)$. Indeed, for any model satisfying (A1) and (A2), the number of ratings required to get at most s misclassified items under SplitSpec has the same scaling in n, w, and q as that of the sample complexity lower bound derived in Theorem 1.

The proof of Theorem 3 constitutes one of the main technical contributions of the paper. It is much more involved than that of existing results on the spectral clustering, e.g. in the Stochastic Block Model (refer to Section 2 for details). The difficulty arises because the entries of the weight matrix of our graph are not strictly independent. SplitSpec is actually designed so as to make this dependency weak enough to be analyzed while preserving most of the information contained in the set of all available ratings.

2. Related Work

Our paper is concerned with recovering clusters from randomly generated data. This topic has attracted a lot of attention recently.

The stochastic block model (SBM) (Holland et al., 1983) may be seen as the simplest way to randomly generate (similarity) data: under this model, the n items are first grouped into K disjoint clusters – the cluster of item i is denoted by $\sigma(i) \in [K]$; the observations are gathered in a symmetric similarity random matrix $A \in \{0,1\}^{n\times n}$ with independent entries, and such that for any i, j, A_{ij} is a Bernoulli r.v. with mean $p_{\sigma(i)\sigma(j)}$. The objective is to recover the clusters solely based on the observation matrix A. The SBM has been heavily studied over the last few years. A first interesting question about the SBM is the *detectability* of the clusters. The latter are detectable if one can devise an algorithm performing better than just randomly assigning items to clusters. The necessary and sufficient condition for detectability (a condition on $p = (p_{k,k'})_{k,k' \in [K]}$) has been identified and established in (Decelle et al., 2011; Mossel et al., 2015b; Massoulié, 2014). Researchers have then investigated conditions under which the clusters can be recovered with a vanishing proportion of misclassified items (when n grows large) or even exactly, see (Yun and Proutiere, 2014; Abbe et al., 2016; Mossel et al., 2015a).

It is worth noting that by choosing $w = 2$ and $q = 1$ in our model, we get a variant of the so-called labeled SBM with sampling. There, an item pair (i, j) is sampled by every user and ratings for these items form a label added to (i, j). The SBM with sampling was studied in (Yun and Proutiere, 2014) and its labeled extension was discussed in (Heimlicher et al., 2012; Yun and Proutiere, 2016). In this paper, we design a spectral algorithm inspired from the spectral method proposed in (Yun and Proutiere, 2016), known to be optimal for the SBM and its extensions. But our paper goes well beyond (Yun and Proutiere, 2016): when $w > 3$, our model clearly departs from the SBM and thus we cannot directly use algorithms designed for the SBM. Instead we propose an algorithm that first creates a random graph between items from the ratings and run a spectral method similar to that used in (Coja-Oghlan, 2010; Yun and Proutiere, 2016) to extract item clusters.

Our model may be seen as a variant of the bipartite stochastic block model (biSBM) discussed in (Feldman et al., 2015) and (Florescu and Perkins, 2016). The biSBM starts with a set $\mathcal{V}$ of n items and a set $\mathcal{U}$ of m users. Both sets are clustered: $\mathcal{V} = \cup_{i=1}^{K} \mathcal{V}_i$ and $\mathcal{U} = \cup_{\ell=1}^{L} \mathcal{U}_\ell$. Edges are generated independently at random between $\mathcal{V}_i$ and $\mathcal{U}_\ell$ with probability $p_{\ell i}$. Hence, the biSBM corresponds to our model with $w = n$ and $q = 1$. Note that in the biSBM, we can make the information sparse by letting $p = (p_{\ell i})_{\ell, i}$ depend on n and m. (Feldman et al., 2015) and (Florescu and Perkins, 2016) studied the symmetric biSBM where $K = L = 2$, $|\mathcal{V}_1| = |\mathcal{V}_2|$, $|\mathcal{U}_1| = |\mathcal{U}_2|$, $p_{11} = p_{22} = \delta p$, and $p_{12} = p_{21} = (2 - \delta)p$. In (Feldman et al., 2015), the authors proposed a subsampled power iteration algorithm extracting the exact clusters when $p = \Omega(\frac{\log(n)}{(\delta-1)^2\sqrt{mn}})$. (Florescu and Perkins, 2016) studied a sharp threshold for the detectability: $p = \Omega\left(\frac{1}{(\delta-1)^2\sqrt{mn}}\right)$ is necessary for detectability and the proposed SBM reduction algorithm detects the item clusters under this condition. They also proposed an algorithm referred to as the diagonal deletion SVD recovering clusters almost exactly when $p = \Omega(\frac{\log(n)}{\sqrt{mn}})$. However, the regime where $\frac{1}{\sqrt{mn}} < p < \frac{\log(n)}{\sqrt{mn}}$ is not treated in these papers. The present paper provides results in a much more general setting, and in particular identifies a necessary and sufficient condition on m or p to get less than s misclassified items. It should also be observed that we could use the SBM reduction algorithm for the case $wq = O(1)$ and the diagonal deletion SVD algorithm when $m = \Omega(\frac{n\log(n)^2}{(wq)^2})$. However, the SBM reduction algorithm becomes inefficient when $wq = \omega(1)$ for it removes too much information and the diagonal deletion SVD algorithm becomes inefficient when $m = O(\frac{n\log(n)^2}{(wq)^2})$ since the input matrix is too noisy. Our algorithm resolves both issues and works well in all cases. Moreover, we show that our algorithm exhibits an optimal (order-wise) sample complexity to guarantee less than s misclassified items (for any given $s = o(n)$).

Finally it is worth mentioning clustering problems under the rich information regime (i.e., when w is large, proportional to n), although the present paper focuses on the sparse information regime. In the rich regime, one can recover the clusters for both items and users, applying algorithms typically found in the collaborative filtering literature. For example, in the model considered in (Aditya et al., 2011; Barman and Dabeer, 2012), there is a simple ground truth rating matrix such that all users in the same cluster give the same rating to all items in the same cluster. The observations consist in a random matrix obtained from the ground truth matrix by erasing entries with probability $1-\epsilon$ and flipping them (i.e., 1 to 0 and 0 to 1) with probability p. This can be seen as a special case of our model when $w = n$, $q = \epsilon$, and $p_{\ell k}$ is p or $1-p$. For $m = n$, the authors of (Xu et al., 2014) provided a tight condition on p and ϵ to recover the exact clusters for both users and items. To conclude, it is worth noting that using matrix completion algorithms, e.g., (Davenport et al., 2014; Candes and Recht, 2012; Keshavan et al., 2010), we can extract the exact average weight matrix when $wq = \Omega(\log(n))$ and $m \geq n$. From there, we can obtain the exact clusters for both users and items. In this paper we can cluster items with much smaller wq, i.e., with very sparse data.

3. Fundamental Performance Limits

In this section, we provide the precise definitions of the inter-cluster divergences used in the lower bounds of the sample complexity derived in Theorems 1 and 2. We further give some important properties of this divergence. Finally we outline the main steps of the proof of Theorem 1 and provide an intuition underlying the proof of Theorem 2. The complete proofs of Theorems 1 and 2 are relegated to the appendix.

3.1. Inter-cluster Divergence

The inter-cluster divergence $D_{ij}(\mathcal{M})$ between clusters i and j represents the hardness of differentiating items from the two clusters. Its definition is motivated by the proof of Theorem 1 outlined below.

For formally defining $D_{ij}(\mathcal{M})$, we need to introduce the following notations. For given $\lambda \leq w$, $\ell \in [L]$, vectors $\boldsymbol{k} = (k_1, ..., k_\lambda) \in [K]^\lambda$ and $\boldsymbol{x} = (x_1, ..., x_\lambda) \in \{0,1\}^\lambda$, we define the function $f_\ell : [K]^\lambda \times \{0,1\}^\lambda \to [0,1]$ as the probability of a type ℓ user to provide the ratings $\boldsymbol{x}$ on λ items whose respective clusters are given by $\boldsymbol{k}$ (i.e., k_t is the cluster of item whose rating is x_t). Formally:

$$f_{\lambda,\ell}(\boldsymbol{k};\boldsymbol{x}) = \prod_{t=1}^{\lambda} p_{\ell k_t}^{x_t}(1-p_{\ell k_t})^{1-x_t} \ .$$

We further define $f_\lambda(\boldsymbol{k};\boldsymbol{x})$ as the weighted sum of $f_{\lambda,\ell}(\boldsymbol{k};\boldsymbol{x})$ with weight β, i.e., $f_\lambda(\boldsymbol{k};\boldsymbol{x}) = \sum_{\ell=1}^{L} \beta_\ell \cdot f_{\lambda,\ell}(\boldsymbol{k};\boldsymbol{x})$. We denote by $f_\lambda(\boldsymbol{k})$ the corresponding probability distribution over $\{0,1\}^\lambda$.

For $i,j \in [K]$ and $\lambda \geq 2$, let $\mathcal{P}_{ij}(\lambda)$ be the set of functions $y_\lambda : [K]^{\lambda-1} \times \{0,1\}^\lambda \to [0,1]$ such that for each $\boldsymbol{k} \in [K]^{\lambda-1}$, $y_\lambda(\boldsymbol{k})$ is a probability distribution over $\{0,1\}^\lambda$ obtained as a convex combination of $f_\lambda(\boldsymbol{k}, i)$ and $f_\lambda(\boldsymbol{k}, j)$. When $\lambda = 1$, $\mathcal{P}_{ij}(\lambda) = \{\gamma f_\lambda(i) + (1-\gamma) f_\lambda(j) : \gamma \in [0,1]\}$.

Next we define the divergence $\Delta_{ij}(\lambda)$ between clusters i and j when users provide exactly λ ratings:

$$\Delta_{ij}(\lambda) = \begin{cases} \min\limits_{\gamma \in [0,1]} \; \max\limits_{i' \in \{i,j\}} \; \mathrm{KL}(\gamma f_1(i) + (1-\gamma) f_1(j) \| f_1(i')) & \text{if } \; \lambda = 1 \\ \min\limits_{y_\lambda \in \mathcal{P}_{ij}(\lambda)} \; \max\limits_{i' \in \{i,j\}} \; \sum\limits_{\boldsymbol{k} \in [K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}\big(y_\lambda(\boldsymbol{k}) \big\| f_\lambda(\boldsymbol{k}, i')\big) & \text{otherwise} \end{cases}$$

where $\alpha_{\boldsymbol{k}} = \prod_{t=1}^{\lambda-1} \alpha_{k_t}$, and where $\mathrm{KL}(a\|b)$ denotes the KL divergence number from distribution a to distribution b. Finally, we are ready to define the divergence between clusters i and j:

$$D_{ij}(\mathcal{M}) = \sum_{\lambda=1}^{w} B(w, \lambda, q) \cdot \Delta_{ij}(\lambda)$$

where $B(w, \lambda, q)$ [1] $:= \binom{w-1}{\lambda-1} \cdot q^{\lambda-1}(1-q)^{w-\lambda}$. The divergence of the model is: $D(\mathcal{M}) = \min_{i \neq j} D_{ij}(\mathcal{M})$.

Next we establish the useful properties of the inter-cluster divergence mentioned in Section 1.

Proposition 4 *Consider $q > 0$. When $w = 1$, $D(\mathcal{M}) = 0$ if and only if there exist $i \neq j \in [K]$ such that $\sum_\ell \beta_\ell (p_{\ell i} - p_{\ell j}) = 0$. When $w \geq 2$, $D(\mathcal{M}) = 0$ if and only if $i \neq j \in [K]$ such that $\forall \ell, p_{\ell i} = p_{\ell j}$.*

Proposition 5 *Assume that $D(\mathcal{M}) > 0$, $(w-1)q \geq \varepsilon > 0$, and that (A1) holds. Then there exists a constant $c(\varepsilon, p) > 0$ depending on $\mathcal{M}$ only through ε and p such that:*

$$c(\varepsilon, p) \leq D(\mathcal{M}) \leq \log(\eta) \; .$$

The proofs of the above propositions can be found in the appendix.

3.2. Proof of Theorem 1

The proof relies on an involved change-of-measure argument that leads to tight lower bounds on the clustering recovery error. This tightness can only be achieved through this kind of argument. We believe that the notion of inter-cluster divergence introduced above provides the exact minimal achievable sample complexity (for given numbers of misclassified items). We outline the proof here, and present a detailed argumentation in the appendix.

Change-of measure. In the following, we refer to Φ, defined by the parameters $\mathcal{M}$, as the true stochastic model under which all the observed ratings are generated, and denote by $\mathbb{P}_\Phi = \mathbb{P}$ (resp. $\mathbb{E}_\Phi[\cdot] = \mathbb{E}[\cdot]$) the corresponding probability measure (resp. expectation). In our change-of-measure argument, we construct a second stochastic model Ψ (whose corresponding probability measure and expectation are $\mathbb{P}_\Psi$ and $\mathbb{E}_\Psi[\cdot]$, respectively). Using a change of measures from $\mathbb{P}_\Phi$ to $\mathbb{P}_\Psi$, we relate the expected number of misclassified items

[1] We let $B(1, 1, q) = 1$ as a consequence of conventions $0^0 = 1$ and $\binom{0}{0} = 1$.

$\mathbb{E}[\varepsilon^\pi(n,m)]$ under any clustering algorithm π to the expected (w.r.t. $\mathbb{P}_\Psi$) log-likelihood ratio $\mathcal{G}$ of the observed ratings under $\mathbb{P}_\Phi$ and $\mathbb{P}_\Psi$. Specifically, we show that, roughly, $\log(n/\mathbb{E}[\varepsilon^\pi(n,m)])$ must be smaller than $\mathbb{E}_\Psi[\mathcal{G}]$ for n large enough.

To construct Ψ, we first pick the clusters i^* and j^* with minimal divergence: $(i^*,j^*) = \arg\min_{i,j:i<j} D_{ij}(\mathcal{M})$. For each $1 \le \lambda \le w$, pick $y^*_\lambda \in \mathcal{P}_{i^*j^*}(\lambda)$ satisfying

$$\Delta_{i^*j^*}(\lambda) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i^*)) = \sum_{k\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, j^*))$$

where such a y^*_λ must exist due to Lemma 10. Using i^*, j^* and y^*_λ, we couple the generation of ratings $\{X_{uv} : u \in \mathcal{U}, v \in \mathcal{V}\}$ under Φ and Ψ in the following way:

C1. The partition of items $\mathcal{V}_1, ..., \mathcal{V}_K$, the set of items assigned to each user $\{\mathcal{W}_u \subset \mathcal{V} : u \in \mathcal{U}\}$, and the set of items that each user gives ratings $\{\mathcal{R}_u \subset \mathcal{W}_u : u \in \mathcal{U}\}$ under Φ are the same as those generated under Ψ.

C2. We select item v^* in $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}$ uniformly at random. If $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} = \emptyset$, we select item v^* in $\mathcal{V}$ uniformly at random. The ratings by user u such that $v^* \notin \mathcal{R}_u$ generated under Ψ are the same as those generated under Φ.

C3. Let $\lambda_u = |\mathcal{R}_u|$. For each user u such that $v^* \in \mathcal{R}_u$, let $\boldsymbol{w}_u := (w_{u,1}, ..., w_{u,\lambda_u-1}, w_{u,\lambda_u} = v^*)$ be the unique sequence of items in $\mathcal{R}_u$ such that $w_{u,t} < w_{u,t+1}$ for all $t < \lambda_u - 1$ and $w_{u,\lambda_u} = v^*$. Regardless of user type, the sequence of ratings $\boldsymbol{x}_u \in \{0,1\}^{\lambda_u}$ on the sequence of items $\boldsymbol{w}_u$ generated under Ψ are observed with probability $y^*_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)$, where $\sigma(\boldsymbol{w}_u) \in [K]^{\lambda_u}$ is the sequence of clusters such that for all $t \le \lambda_u$, $\sigma(w_{u,t}) = k$ if $w_{u,t} \in \mathcal{V}_k$.

Log-likelihood ratio and its connection to the number of misclassified items. We introduce the log-likelihood ratio of the observed ratings $\{x_{uv} : u \in \mathcal{U}, v \in \mathcal{V}\}$ under Ψ and Φ as:

$$\mathcal{G} = \sum_{u\in\mathcal{U}} \mathbb{1}_{[v^*\in\mathcal{R}_u]} \log \frac{y^*_{\lambda_u}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_{\lambda_u}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} = \sum_{u\in\mathcal{U}} \sum_{\lambda=1}^{w} \mathbb{1}_{[v^*\in\mathcal{R}_u, \lambda_u=\lambda]} \log \frac{y^*_{\lambda}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_{\lambda}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}$$

where $\mathbb{1}_{\mathcal{C}}$ is the indicator of event $\mathcal{C}$ and $\boldsymbol{w}_u$, $\sigma(\boldsymbol{w}_u)$, and $\boldsymbol{x}_u$ are defined in C3. Let π be a clustering algorithm that outputs $\hat{\mathcal{V}}_1, ..., \hat{\mathcal{V}}_K$ such that $\sum_{k=1}^K \hat{\mathcal{V}}_k = \mathcal{V}$ and $\hat{\mathcal{V}}_i \cap \hat{\mathcal{V}}_j = \emptyset$ for all $i \neq j$. Without loss of generality, we assume $\left|\bigcup_{k=1}^K \hat{\mathcal{V}}_k \setminus \mathcal{V}_k\right| \le \left|\bigcup_{k=1}^K \hat{\mathcal{V}}_{\theta(k)} \setminus \mathcal{V}_k\right|$ for any permutation θ of $[K]$. Let $\mathcal{E}$ be the set of misclassified items by π, i.e., $\mathcal{E} = \bigcup_{k=1}^K \hat{\mathcal{V}}_k \setminus \mathcal{V}_k$. Then we have $|\mathcal{E}| = \varepsilon^\pi(n,m)$.

Next we establish a connection between the number of misclassified items and the distribution of $\mathcal{G}$ under Ψ. Formally, if the algorithm π satisfies $\mathbb{E}[\varepsilon^\pi(n,m)] \le s$, then

$$\log(n/s) - \log(2/\alpha_{i^*}) \le \mathbb{E}_\Psi\left[\mathcal{G}\right] + \sqrt{\frac{3}{\alpha_{i^*}} \mathbb{E}_\Psi[(\mathcal{G} - \mathbb{E}_\Psi[\mathcal{G}])^2]}\,.$$

Analysis of the log-likelihood ration $\mathcal{G}$. To complete the proof, we derive an upper bound of the r.h.s. of the above inequality. Specifically, we prove that $\mathbb{E}[\mathcal{G}] = \frac{mwp}{n} D(\mathcal{M})$ by definition of i^*, j^* and $D(\mathcal{M})$. Furthermore, when $wq = o(\sqrt{n})$, then $\sqrt{\mathbb{E}_\Psi[(\mathcal{G} - \mathbb{E}_\Psi[\mathcal{G}])^2]} = o(\frac{mwp}{n} D(\mathcal{M}))$, which completes the proof.

3.3. Proof of Theorem 2

For the proof of Theorem 2, we also use a change-of-measure argument. However, we connect the sample complexity to an upper bound $\tilde{D}_{ij}(\mathcal{M})$ of the inter-cluster divergence $D_{ij}(\mathcal{M})$. To investigate an item v's cluster more efficiently, an adaptive assignment can exploit the previously collected ratings. Hence, to obtain a lower bound of the sample complexity under the adaptive assignment scenario, we use $\tilde{D}_{ij}(\mathcal{M})$ with the *maximal* KL divergence $\tilde{\Delta}_{ij}(\lambda)$ between clusters i and j instead of the *average* KL divergence $\Delta_{ij}(\lambda)$, where the maximum and average are taken over every configuration $\boldsymbol{k} \in [K]^{\lambda-1}$ of the assigned items' clusters. Formally,

$$\tilde{D}_{ij}(\mathcal{M}) := \sum_{\lambda=1}^{w} B(w, \lambda, q) \cdot \tilde{\Delta}_{ij}(\lambda) \qquad \text{and,}$$

$$\tilde{\Delta}_{ij}(\lambda) := \begin{cases} \max\left\{\mathrm{KL}(f_1(i)\|f_1(j)), \mathrm{KL}(f_1(j)\|f_1(i))\right\} & \text{if } \lambda = 1 \\ \max\limits_{\boldsymbol{k}\in[K]^{\lambda-1}} \max\left\{\mathrm{KL}(f_\lambda(\boldsymbol{k}, i)\|f_\lambda(\boldsymbol{k}, j)), \mathrm{KL}(f_\lambda(\boldsymbol{k}, j)\|f_\lambda(\boldsymbol{k}, i))\right\} & \text{otherwise} \end{cases}.$$

We provide the formal proof of Theorem 2 in the appendix, where we establish a precise connection between $\tilde{D}(\mathcal{M})$ and the sample complexity using a change-of-measure argument. We note that we have $D(\mathcal{M}) \leq \tilde{D}(\mathcal{M}) \leq \frac{\tilde{D}(\mathcal{M})}{(\alpha_1/3)}$ due to the definition of $\tilde{D}(\mathcal{M})$. This implies that comparing to random assignments, adaptively selecting items to rate can reduce the sample complexity by a constant factor, while the asymptotic order of the sample complexity remains the same, i.e., $mwq = \Omega(n \log(n/s))$ regardless of the assignment scheme.

4. The SplitSpec Algorithm

This section presents SplitSpec, a spectral clustering algorithm that recovers the clusters from the available ratings. The algorithm consists of two steps. In the first and main step, SplitSpec constructs a weighted graph whose vertices are items, and whose edges and corresponding weights are constructed from the ratings. The second step applies the spectral method to the obtained weighted graph to recover the clusters.

4.1. Graph Construction (the Split step)

A simple way to generate edges would be to draw an edge between items v and v' if there exists a user u that rated v and v' positively, and all other items in $\mathcal{R}_u$ negatively. The edges obtained that way would be independent and the corresponding graph would be the same as in a classical stochastic block model (Yun and Proutiere, 2014). However, this construction would not exploit all the information available from the ratings, especially when $wq = \omega(1)$.

To circumvent this difficulty, for each user u, we randomly split $\mathcal{W}_u$ into several disjoint groups so that a constant fraction of information can be kept. We consider each group separately, and a group will generate an edge between two items if and only if these items are the only two positively rated items within the group. Now the groups are built so that with positive probability, each of them generates an edge. More precisely, the objective is that the cardinality of groups is selected so that the probability of having only two positively

Algorithm 1 The Split step – Generating the weight matrix A

```
Input: Observation matrix $X \in \mathbb{N}^{N\times N}$.
Rating intensity: $\tilde{q} \leftarrow \frac{\sum_{v=1}^{n}\sum_{u=1}^{m} X_{uv}}{mw}$
Group size: $\gamma \leftarrow w \wedge \lfloor 2/\tilde{q} \rfloor$
Number of groups: $h \leftarrow \lfloor w/\gamma \rfloor$
Initialization: $A \leftarrow \mathbf{0} \in \mathbb{R}^{N\times N}$
for $u = 1$ to $m$ do
  $B_1, \dots, B_h \leftarrow$ random subsets of $\mathcal{W}_u$ such that $|B_k| = \gamma$ for all $k$ and $B_k \cap B_j = \emptyset$ for all $k \neq j$
  (randomly assigning items in $\mathcal{W}_u$ to groups)
  for $k = 1$ to $h$ do
    $A_{vv'} \leftarrow A_{vv'} + 1$ when $\sum_{v\in B_k} X_{uv} = 2$, $\{v, v'\} \subset B_k$, and $X_{uv} = X_{uv'} = 1$
  end for
end for
Output: $A$.
```

rated items in a group is bounded away from zero. To meet this objective, if $\tilde{q}$ denotes the probability of an item to be positively rated, then we can choose groups of cardinality[2] $\lfloor 2/\tilde{q} \rfloor$. Note that $\tilde{q}$ can be accurately estimated by $\frac{1}{mw}(\sum_u \sum_v X_{uv})$ where X_{uv} is a binary variable equal to 1 iff user u rated item v positively. Hence we first estimate $\tilde{q}$ and then randomly divide $\mathcal{W}_u$ into $h = \lfloor \frac{w}{\gamma} \rfloor$ groups of equal size $\gamma = w \wedge \lfloor 2/\tilde{q} \rfloor$[3] (if γ does not divide w, $\mathcal{W}_u$ is not fully covered by the groups). If $w\tilde{q} \geq 2$, then one can check that $\Theta(w\tilde{q})$ edges are generated from the groups associated to $\mathcal{W}_u$. If $w\tilde{q} < 2$, $\mathcal{W}_u$ consists of a single group, and an edge is generated with probability $\Theta(w^2\tilde{q}^2)$.

Now in the weighted graph constructed through the above procedure, each time an edge between items v and v' is generated, the weight of (v, v') is incremented by 1. We denote by $A \in \mathbb{R}^{n\times n}$ the weight matrix of the graph. The construction of A is performed as described in Algorithm 1.

4.2. Spectral Partition (the Spec step)

The second step of the SplitSpec algorithm consists in using a spectral partition algorithm similar to that used in (Yun and Proutiere, 2016). The partition is applied to the matrix A to recover the clusters, and its pseudo-code is presented in Algorithm 2. It should be noted that we cannot directly use the theoretical results derived in (Yun and Proutiere, 2016) since the edges generated in Algorithm 1 are not independent. Indeed, at most γ edges can be generated by a user and the edges generated by the same user cannot share any item, and hence are not independent. This is the main technical difficulty for the analysis: we will show that when $wq = o(\sqrt{n})$, we can overcome the statistical dependence in the entries of A.

The Spec step consists of three parts.

[2] Indeed, if $Z_1, \dots, Z_\gamma$ are i.i.d. Bernoulli r.v. with $\mathbb{P}[Z_i = 1] = \tilde{q}$ and $\gamma = \lfloor 2/\tilde{q} \rfloor$, then $\mathbb{P}[\sum_{i=1}^{\gamma} Z_i = 2] = \Theta(1)$.

[3] We use standard notation $\wedge$ to denote min so that $x \wedge y = \min\{x, y\}$.

Algorithm 2 The Spec step – Spectral Partition

```
Input: Weighted matrix $A \in \mathbb{R}^{N \times N}$ and cluster number $K$.
Estimated average degree: $\tilde{p} \leftarrow \frac{\sum_{v,v'} A_{vv'}}{n(n-1)}$
1. Trimming process. Construct $A_\Gamma = (A_{vv'})_{v,v' \in \Gamma}$, where $\Gamma$ is the set of items obtained after removing $\lfloor n \exp(-n\tilde{p}) \rfloor$ items having the largest $\sum_{v' \in V} A_{vv'}$.
2. Spectral decomposition. Run Algorithm 3 with input $A_\Gamma, \tilde{p}$, and output $(\mathcal{S}_k)_{k=1,\dots,K}$.
3. Successive improvement.
$\hat{p}(i,j) \leftarrow \frac{\sum_{v \in \mathcal{S}_i} \sum_{v' \in \mathcal{S}_j} A_{vv'}}{|\mathcal{S}_i| m}$ for all $1 \le i, j \le K$
$\mathcal{S}_k^{(0)} \leftarrow \mathcal{S}_k$ for all $1 \le k \le K$
for $t = 1$ to $\log n$ do
  $\mathcal{S}_k^{(t)} \leftarrow \emptyset$ for all $1 \le k \le K$
  for $v \in \mathcal{V}$ do
    $k^* = \arg\max_{1 \le k \le \hat{K}} \left\{ m \log(1 - \sum_{i=1}^K \hat{p}(k,i)) + \sum_{i=1}^K \sum_{w \in \mathcal{S}_i^{(t-1)}} A_{vw} \log \frac{\hat{p}(k,i)}{1 - \sum_{i=1}^K \hat{p}(k,i)} \right\}$
    (tie broken uniformly at random)
    $\mathcal{S}_{k^*}^{(t)} \leftarrow \mathcal{S}_{k^*}^{(t)} \cup \{v\}$
  end for
end for
$\hat{\mathcal{V}}_k \leftarrow \mathcal{S}_k^{(\log n)}$ for all $1 \le k \le K$
Output: $(\hat{\mathcal{V}}_k)_{k=1,\dots,K}$.
```

1. **Trimming process.** In the first part, we trim the entries of A so that the remaining matrix enjoys some regularity property. More precisely, the trimming step removes the $n \exp(-H)$ items v having the largest sum of positive entries (i.e. $\sum_{v'=1}^n A_{vv'}$) where $H = \frac{1}{n} \sum_{v=1}^n \sum_{v'=1}^n A_{vv'}$. The remaining items do not have too many positive entries. Let A_Γ be the output of the trimming step.

2. **Spectral decomposition.** In the second part, we provide initial estimates of the clusters by applying Algorithm 3, presented in the appendix, to the trimmed matrix A_Γ. Note that $\mathbb{E}[A_\Gamma]$ is of rank K and has a block structure, with blocks corresponding to the true clusters. Hence the spectral analysis of its noisy version A_Γ can provide good estimates of the clusters. The trimming step ensures that the spectral norm of the noise matrix $A_\Gamma - \mathbb{E}[A_\Gamma]$ is small and that the rank K approximation of A_Γ accurately approximates $\mathbb{E}[A_\Gamma]$. Algorithm 3 and its analysis are classical (Coja-Oghlan, 2010), (Yun and Proutiere, 2016), and in turn, we will show that the number of misclassified items after this algorithm is $O(n/H)$. In regimes of interest, we will further establish that $n/H = o(n)$, so that already after the second part of the Spec step, we have very good estimates of the clusters. This statement is made precise in Theorem 6 below.

3. **Successive improvements.** Using the already accurate estimates of the clusters, we can estimate $\hat{p}(i,j)$ the probability that a user generates an edge between an item of cluster $\mathcal{V}_i$ and items in cluster $\mathcal{V}_j$. From $\hat{p} = (\hat{p}(i,j))_{i,j \in [K]}$, if correctly estimated, we

can compute for an item v the log-likelihood of the observations given that v belongs to a given cluster $\mathcal{V}_k$: $m\log(1-\sum_{i=1}^K \hat{p}(k,i)) + \sum_{i=1}^K \sum_{v'\in\mathcal{V}_i} A_{vv'} \log \frac{\hat{p}(k,i)}{1-\sum_{i=1}^K \hat{p}(k,i)}$. In the successive improvements, we just assign item v to the cluster maximizing this likelihood. The analysis of these improvements is summarized in Theorem 7.

As mentioned above, the spectral decomposition yields cluster estimates $\mathcal{S}_1,\dots,\mathcal{S}_K$ satisfying: with high probability,

$$\min_\theta \sum_{k=1}^K |\mathcal{S}_{\theta(k)} \setminus \mathcal{V}_k| = O\left(\frac{n}{H}\right), \tag{1}$$

where the minimum is taken w.r.t. permutation θ of $[K]$. The analysis of the scaling of H can be done as follows. In Algorithm 1, for each user, we create h disjoint groups of cardinalities γ. Each group randomly generates at most an edge with probability $\Theta\left(\gamma^2 q^2\right)$ (since the probability that the user rates two items out of the γ items in the group scales as $\gamma^2 q^2$). Hence, the algorithm generates $\Theta\left(mh\gamma^2 q^2\right)$ edges. Since $\gamma = w \wedge \lfloor 2/\tilde{q} \rfloor$, $h = \lfloor w/\gamma \rfloor$, and $\tilde{q}$ exhibits the same scaling as q, we have $h\gamma^2 q^2 = \Theta(wq(1\wedge wq))$ and thus

$$H = \Theta\left(\frac{mwq(1\wedge wq)}{n}\right). \tag{2}$$

The following theorem formalizes and combines (1) and (2). Its proof is presented in the appendix.

Theorem 6 *Assume that $D(\mathcal{M}) > 0$, $h = o(\sqrt{n})$ and $h^2 m = o(n^2)$ or equivalently $wq = o(\sqrt{n})$ and $(wq)^2 m = o(n^2)$. After the spectral decomposition in Algorithm 2, with high probability, we have*

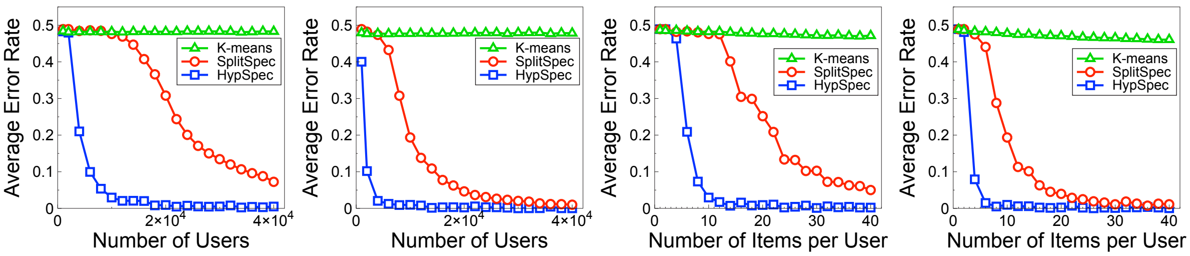
$$\min_\theta \left| \bigcup_{k=1}^K \mathcal{S}_{\theta(k)} \setminus \mathcal{V}_k \right| = O\left(\frac{n^2}{mwq(1\wedge wq)}\right).$$

Using the improvement steps after the spectral decomposition allows us to refine the cluster estimates. The next theorem quantifies the number of ratings one should observe so that the SplitSpec algorithm misclassifies s items at most. Note that this number has the same scaling in n, w, q, and s as our sample complexity lower bound. Thus, SplitSpec is order-optimal.

Theorem 7 *When $D(\mathcal{M}) > 0$, $h = o(\sqrt{n})$ and $h^2 m = o(n^2)$ or equivalently $wq = o(\sqrt{n})$ and $(wq)^2 m = o(n^2)$, there exists a constant $C > 0$ such that when*

$$mwq \geq \frac{Cn\log(n/s)}{(1\wedge wq)},$$

with high probability, the number of misclassified items under SplitSpec *is less than $s = o(n)$.*



(a) $q=0.5, w=10$ (b) $q=1, w=10$ (c) $q=0.5, m=10,000$ (d) $q=1, m=10,000$

Figure 1: The average error rate of various algorithms clustering $1,000$ items from ratings from the collaborative clustering model $\mathcal{M} = (\alpha, \beta, p, w, q)$ with $\alpha = (0.5, 0.5)$, $\beta = (0.5, 0.5)$, $p = (0.8, 0.2; 0.2, 0.8)$, $q = 0.5$ or 1; (a)-(b) $w = 10$ and varying m; (c)-(d) $m = 10,000$ and varying w.

5. Numerical Experiments

Next we briefly present experimental results supporting our analytical findings and illustrating the performance of SplitSpec on synthetic datasets.

Algorithms. We compare three algorithms: K-means, SplitSpec, and HypSpec. For the first algorithm, we represent each item by the proportion of positive ratings this item has received, and we simply run k-means algorithm (Lloyd, 1982) on this one-dimensional data (we use random initialization). For the last algorithm, HypSpec, we first generate a weight matrix $\tilde{A}$ such that $\tilde{A}_{vv'}$ is the number of users who give $+1$'s on both of items v and v'. Then, we run Algorithm 2 with input $\tilde{A}$.

When a user positively rates a subset S of items, to keep that information fully, we would need to add the item graph an hyper-edge consisting of this set S. This is what is done in HypSpec by constructing $\tilde{A}$. In the construction of A in SplitSpec, we discard part of the hyper-edge information, but the spectral analysis A becomes possible because its entries are by construction *almost* independent. On the contrary, the entries of $\tilde{A}$ exhibit a strong dependency, and it is very unlikely that the performance of HypSpec can be analyzed.

Data. We consider $1,000$ items having two clusters with equal sizes ($\alpha_1 = \alpha_2 = 0.5$). We randomly generates ratings from two types of users ($\beta_1 = \beta_2 = 0.5$). Type-1 user rating statistics are given by $(p_{11}, p_{12}) = (0.8, 0.2)$, and those of type-2 users by $(p_{21}, p_{22}) = (0.2, 0.8)$. Note that the average (over users) rating of an item does not depend on the cluster of this item. Hence, K-means cannot separate the two clusters, since it does not exploit users' structure.

Results. Figure 1 presents our results. In Figures 1a-1b, we fix w and q and vary the number of users, whereas in Figures 1c-1d, the number of users is fixed and w varies. Each error rate reported in the curves are averaged of 200 realizations of the random ratings. As expected K-means does not do better than just randomly assigning items to clusters. Naturally the error rates of HypSpec and SplitSpec decrease with the number of ratings mwq. Note that HypSpec outperforms SplitSpec, but as explained above, HypSpec is challenging to analyze and hence has no theoretical performance guarantees.

6. Conclusion

In this paper, we have studied the collaborative clustering problem whose objective is to recover a hidden cluster structure among items from random ratings provided by heterogeneous users. We have first derived fundamental performance limits for this problem: the number of ratings required to guarantee the existence of an algorithm estimating the clusters with a prescribed level of accuracy depends on the notion of inter-cluster divergence. We have then proposed SplitSpec, an algorithm that achieves these fundamental performance limits order-wise. We believe that our lower bound on the sample complexity is tight, and will explore extensions of our algorithms and its analysis to establish this tightness formally. SplitSpec was actually designed so that its spectral step can be analyzed using techniques from random matrix theory, but this required that the algorithm does not leverage all the information provided by the users' ratings. To enhance the performance of SplitSpec, we need to modify the algorithm so as it can fully exploit this information. This is a promising research direction as illustrated in our numerical experiments, where we tested HypSpec, an algorithm that fully exploits users' ratings, but that appears really hard to analyze.

References

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

S. T. Aditya, O. Dabeer, and B. K. Dey. A channel coding perspective of collaborative filtering. *IEEE Transactions on Information Theory*, 57(4):2327–2341, 2011.

O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Fast and provably good seedings for k-means. In *Proc. of NIPS*, 2016.

K. Barman and O. Dabeer. Analysis of a collaborative filter based on popularity amongst neighbors. *IEEE Transactions on Information Theory*, 58(12):7110–7134, 2012.

E. Candes and B. Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

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

M. D. Ekstrand, J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, 2011.

U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

V. Feldman, W. Perkins, and S. Vempala. Subsampled power iteration: a unified algorithm for block models and planted CSP's. In *Proc. of NIPS*, 2015.

L. Florescu and W. Perkins. Spectral thresholds in the bipartite stochastic block model. In *Proc. of COLT*, 2016.

S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.

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

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proc. of STOC*, 2014.

E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proc. of STOC*, 2015a.

E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015b.

A. B. Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 French original. translated by V Zaiats, 2009.

J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying. Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs. In *Proc. of ACM Sigmetrics*, 2014.

Y. Yue, C. Wang, K. El-Arini, and C. Guestrin. Personalized collaborative clustering. In *Proc. of WWW*, 2014.

S.-Y. Yun and A. Proutiere. Community detection via random and adaptive sampling. In *Proc. of COLT*, 2014.

S.-Y. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Proc. of NIPS*, 2016.

Appendix A. Proof of Proposition 4

We start with $w = 1$. Note that $\sum_{\ell=1}^{L} \beta_\ell p_{\ell i} = \sum_{\ell=1}^{L} \beta_\ell p_{\ell j}$ if and only if $\Delta_{ij}(1) = 0$ since the KL divergence of two Bernoulli distributions is 0 only if their parameters are identical. Note that when $w = 1$, $D_{ij}(\mathcal{M}) = \Delta_{ij}(1)$. Hence, it follows that when $w = 1$, $D(\mathcal{M}) = \min_{i \neq j} \Delta_{ij}(1) = 0$ if and only if there exist $i \neq j \in [K]$ such that $\sum_{\ell=1}^{L} \beta_\ell p_{\ell i} = \sum_{\ell=1}^{L} \beta_\ell p_{\ell j}$.

We now focus on $w \geq 2$. When there exist $i' \neq j' \in [K]$ such that $p_{\ell i'} = p_{\ell j'}$ for all $\ell \in [L]$, then $\Delta_{i'j'}(\lambda) = 0$ for all $\lambda \leq w$ which implies $D_{i'j'}(\mathcal{M}) = 0$ and $D(\mathcal{M}) = \min_{i \neq j} D_{ij}(\mathcal{M}) = 0$ since $D(\mathcal{M})$ is non-negative. Conversely, we will show that if $D(\mathcal{M}) = 0$, then there exist $i' \neq j' \in [K]$ such that $p_{\ell i'} = p_{\ell j'}$ for all $\ell \in [L]$. To this end, we use two useful lemmas on $\Delta_{ij}(\lambda)$ in the followings:

Lemma 8 *For given $i \neq j \in [K]$ and $\lambda \geq 1$, we have*

$$\Delta_{ij}(\lambda + 1) \;\geq\; \Delta_{ij}(\lambda) \;.$$

Lemma 9 *For any given $i \neq j \in [K]$ and $\lambda \geq 2$, we have*

$$\Delta_{ij}(\lambda) \;\geq\; \alpha_1 \left(\sum_{\ell=1}^{L} \beta_\ell (p_{\ell i} - p_{\ell j})^2 \right)^2 \;.$$

We provide proofs of Lemmas 8 and 9 in Appendices A.1 and A.2, respectively. From Lemma 8, it follows that

$$D(\mathcal{M}) \;\geq\; \min_{i \neq j \in [K]} (1 - (1-q)^{w-1}) \cdot \Delta_{ij}(2) \tag{3}$$

since $\sum_{\lambda=1}^{w} B(w, \lambda, q) = 1$ and $B(w, 1, q) = (1-q)^{w-1}$. Thus, Lemma 9 with (3) implies that $D(\mathcal{M}) = 0$ only if there exist $i' \neq j' \in [K]$ such that $\Delta_{i'j'}(2) = 0$, where $\Delta_{i'j'}(2) = 0$ only if $p_{\ell i'} = p_{\ell j'}$ for all $\ell \in [L]$. This completes the proof of Proposition 4.

A.1. Proof of Lemma 8

Pick $y_{\lambda+1} \in \mathcal{P}_{ij}(\lambda+1)$. By the definition of $\mathcal{P}_{ij}(\lambda+1)$, for each $\boldsymbol{k} \in [K]^\lambda$, we can find a constant $\gamma(\boldsymbol{k}) \in [0,1]$ such that

$$y_{\lambda+1}(\boldsymbol{k}) = \gamma(\boldsymbol{k}) f_{\lambda+1}(\boldsymbol{k}, i) + (1 + \gamma(\boldsymbol{k})) f_{\lambda+1}(\boldsymbol{k}, j) \;.$$

Then we construct $y_\lambda : [K]^{\lambda-1} \times \{0,1\}^\lambda \to [0,1]$ by marginalizing out $k' \in [K]$ and $x' \in \{0,1\}$ from $y_{\lambda+1}(k', \boldsymbol{k}; x', \boldsymbol{x})$ for each $\boldsymbol{k} \in [K]^{\lambda-1}$ and $\boldsymbol{x} \in \{0,1\}^\lambda$. Formally,

$$\begin{aligned} y_\lambda(\boldsymbol{k}; \boldsymbol{x}) &:= \sum_{k' \in [K]} \alpha_{k'} \sum_{x' \in \{0,1\}} y_{\lambda+1}(k', \boldsymbol{k}; x', \boldsymbol{x}) \\ &= \sum_{k' \in [K]} \alpha_{k'} \sum_{x' \in \{0,1\}} \gamma(k', \boldsymbol{k}) f_{\lambda+1}(k', \boldsymbol{k}, i; x', \boldsymbol{x}) + (1 + \gamma(k', \boldsymbol{k})) f_{\lambda+1}(k', \boldsymbol{k}, j; x', \boldsymbol{x}) \\ &= \sum_{k' \in [K]} \alpha_{k'} \left(\gamma(k', \boldsymbol{k}) f_\lambda(\boldsymbol{k}, i; \boldsymbol{x}) + (1 + \gamma(k', \boldsymbol{k})) f_\lambda(\boldsymbol{k}, j; \boldsymbol{x}) \right) \;, \end{aligned}$$

where for the last inequality, we use the fact that for each ℓ, $f_{\lambda+1,\ell}(k', \boldsymbol{k}, i; x', \boldsymbol{x}) = (p_{\ell,k'})^{x'} (1 - p_{\ell,k'})^{1-x'} f_{\lambda,\ell}(\boldsymbol{k}, i; \boldsymbol{x})$. We further have

$$y_\lambda(\boldsymbol{k}; \boldsymbol{x}) = \left(\sum_{k' \in [K]} \alpha_{k'} \gamma(k', \boldsymbol{k}) \right) \cdot f_\lambda(\boldsymbol{k}, i; \boldsymbol{x}) + \left(\sum_{k' \in [K]} \alpha_{k'} (1 - \gamma(k', \boldsymbol{k})) \right) \cdot f_\lambda(\boldsymbol{k}, j; \boldsymbol{x}) \;,$$

which implies that for each $\boldsymbol{k} \in [K]^{\lambda-1}$, $y_\lambda(\boldsymbol{k})$ is a convex combination of $f_\lambda(\boldsymbol{k}, i)$ and $f_\lambda(\boldsymbol{k}, j)$, i.e., $y_\lambda \in \mathcal{P}_{ij}(\lambda, p)$. Using the log-sum inequality or the convexity of KL divergence, it is not hard to check that for $\boldsymbol{k} \in [K]^{\lambda-1}$,

$$\begin{aligned}\sum_{k' \in [K]} \alpha_{k'} \mathrm{KL}\big(y_{\lambda+1}(k', \boldsymbol{k}) \big\| f_{\lambda+1}(k', \boldsymbol{k}, i)\big) &\geq \mathrm{KL}\left(\sum_{k' \in [K]} \alpha_{k'} y_{\lambda+1}(k', \boldsymbol{k}) \middle\| \sum_{k' \in [K]} \alpha_{k'} f_{\lambda+1}(k', \boldsymbol{k}, i)\right) \\ &\geq \mathrm{KL}(y_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i)) \ ,\end{aligned}$$

where for the first inequality, we use the convexity of KL divergence, and for the last inequality, we use the construction of y_λ, the log-sum inequality and the fact that for each $\boldsymbol{x} \in \{0,1\}^\lambda$ $f_{\lambda+1}(k', \boldsymbol{k}, i; \boldsymbol{x}) = \sum_{k' \in [K]} \alpha_{k'} \sum_{x' \in \{0,1\}} f_{\lambda+1}(k', \boldsymbol{k}, i; x', \boldsymbol{x})$. This completes the proof of Lemma 8.

A.2. Proof of Lemma 9

Due to Lemma 8, it is enough to show the lemma with $\lambda = 2$. Pick $y \in \mathcal{P}_{ij}(2)$ such that

$$\Delta_{ij}(2) = \sum_{k \in [K]} \alpha_k \cdot \mathrm{KL}(y(k) \| f_\lambda(k, i)) = \sum_{k \in [K]} \alpha_k \cdot \mathrm{KL}(y(k) \| f_\lambda(k, j)), \tag{4}$$

where such a y must exist due to Lemma 10. Then,

$$\begin{aligned}\Delta_{ij}(2) &= \frac{1}{2} \sum_{k \in [K]} \alpha_k \cdot (\mathrm{KL}(y(k) \| f_2(k, i)) + \mathrm{KL}(y(k) \| f_2(k, j))) \\ &= -\sum_{k \in [K]} \alpha_k \cdot \sum_{x, x' \in \{0,1\}} y(k; x, x') \log\left(\frac{\sqrt{f_2(k, i; x, x') f_2(k, j; x, x')}}{y(k; x, x')}\right) \\ &\geq \sum_{k \in [K]} \alpha_k \cdot \sum_{x, x' \in \{0,1\}} y(k; x, x') - \sqrt{f_2(k, i; x, x') f_2(k, j; x, x')} \\ &= \frac{1}{2} \sum_{k \in [K]} \alpha_k \cdot \sum_{x, x' \in \{0,1\}} \left(\sqrt{f_2(k, i; x, x')} - \sqrt{f_2(k, j; x, x')}\right)^2 \\ &\geq \frac{1}{8} \sum_{k \in [K]} \alpha_k \cdot \left(\sum_{x, x' \in \{0,1\}} \big|f_2(k, i; x, x') - f_2(k, j; x, x')\big|\right)^2\end{aligned}$$

where for the last inequality, we use Le Cams inequality. Noting that $\alpha_1 \leq ... \leq \alpha_K$, we further have

$$\begin{aligned}&\Delta_{ij}(2) \\ &\geq \frac{1}{8} \sum_{k \in [K]} \alpha_k \cdot \left(\sum_{x, x' \in \{0,1\}} \big|f_2(k, i; x, x') - f_2(k, j; x, x')\big|\right)^2 \\ &\geq \frac{1}{8} \alpha_1 \cdot \left(\sum_{x, x' \in \{0,1\}} \big|f_2(i, i; x, x') - f_2(i, j; x, x')\big|\right)^2 + \frac{1}{8} \alpha_1 \cdot \left(\sum_{x, x' \in \{0,1\}} \big|f_2(j, i; x, x') - f_2(j, j; x, x')\big|\right)^2\end{aligned}$$

$$\geq \frac{1}{16}\alpha_1 \cdot \Bigg(\sum_{x,x' \in \{0,1\}} \big|f_2(i,i;x,x') - f_2(i,j;x,x')\big| + \big|f_2(j,i;x,x') - f_2(j,j;x,x')\big| \Bigg)^2 \tag{5}$$

where for the last inequality, we use the fact that $x^2 + y^2 \geq \frac{(x+y)^2}{2}$. From the definition of f_2, it follows that

$$\begin{aligned}
&\frac{1}{2} \sum_{x,x' \in \{0,1\}} \big|f_2(i,i;x,x') - f_2(i,j;x,x')\big| + \big|f_2(j,i;x,x') - f_2(j,j;x,x')\big| \\
&\geq \Bigg|\sum_{\ell \in [L]} \beta_\ell p_{\ell i}(p_{\ell i} - p_{\ell j})\Bigg| + \Bigg|\sum_{\ell \in [L]} \beta_\ell p_{\ell j}(p_{\ell j} - p_{\ell i})\Bigg| \\
&\quad + \Bigg|\sum_{\ell \in [L]} \beta_\ell (1 - p_{\ell i})\big((1 - p_{\ell i}) - (1 - p_{\ell j})\big)\Bigg| + \Bigg|\sum_{\ell \in [L]} \beta_\ell (1 - p_{\ell j})\big((1 - p_{\ell j}) - (1 - p_{\ell i})\big)\Bigg| \\
&\geq \Bigg|\sum_{\ell \in [L]} \beta_\ell (p_{\ell i} - p_{\ell j})^2\Bigg| + \Bigg|\sum_{\ell \in [L]} \beta_\ell \big((1 - p_{\ell i}) - (1 - p_{\ell j})\big)^2\Bigg| \\
&\geq 2\Bigg|\sum_{\ell \in [L]} \beta_\ell (p_{\ell i} - p_{\ell j})^2\Bigg|
\end{aligned}$$

which completes the proof of Lemma 9 with (5).

Appendix B. Proof of Proposition 5

Using (A1), it is straightforward to check that for $i, j \in [K]$, $\gamma \in [0,1]$, $\boldsymbol{k} \in [K]^{\lambda-1}$, and $\boldsymbol{x} \in \{0,1\}^\lambda$,

$$\frac{\gamma f_\lambda(\boldsymbol{k}, i; \boldsymbol{x}) + (1-\gamma) f_\lambda(\boldsymbol{k}, j; \boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i; \boldsymbol{x})} \;\leq\; \frac{\gamma f_\lambda(\boldsymbol{k}, i; \boldsymbol{x}) + \eta \cdot (1-\gamma) f_\lambda(\boldsymbol{k}, i; \boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i; \boldsymbol{x})} \;\leq\; \eta \tag{6}$$

which implies $\Delta_{ij}(\lambda) \leq \log \eta$ and thus $D(\mathcal{M}) \leq \log \eta$ since $\sum_{\lambda=1}^{w} B(w, \lambda, q) = 1$.

We now obtain the lower bound of $D(\mathcal{M})$. From Proposition 4, there exists $\varepsilon' > 0$ such that for some $i \neq j \in [K]$ and $\ell \in [L]$, $(p_{\ell i} - p_{\ell j})^2 \geq \varepsilon'$ since we assume $D(\mathcal{M}) > 0$ and $(w-1)q \geq \varepsilon > 0$, i.e., $w \geq 2$. Noting $\sum_{\lambda=1}^{w} B(w, \lambda, q) = 1$ and $B(w, 1, q) = (1-q)^{w-1}$, we have

$$\begin{aligned}
D(\mathcal{M}) \;\geq\; (1 - (1-q)^{w-1}) \cdot \Delta_{ij}(\lambda) \;&\geq\; \frac{q(w-1)}{1 + q(w-1)} \cdot \Delta_{ij}(\lambda) \\
&\geq\; \frac{\varepsilon}{1+\varepsilon} \cdot \Delta_{ij}(\lambda) \\
&\geq\; \frac{\varepsilon}{1+\varepsilon} \cdot \alpha_1 \left(\beta_\ell (p_{\ell i} - p_{\ell j})^2\right)^2
\end{aligned}$$

where for the last inequality, we use Lemma 9. Since $(p_{\ell i} - p_{\ell j})^2 \geq \varepsilon' > 0$, this completes the proof of Proposition 5.

Appendix C. Proof of Theorem 1

We will use a similar change-of-measure argument used for Theorem 1 in (Yun and Proutiere, 2016). In the following, we refer to Φ as the true stochastic model of ratings with the random assignment. We first construct a slightly perturbed model Ψ coupled with Φ.

Construction of Ψ. We couple the generation of ratings under Φ and Ψ as follows. Let $(i^*, j^*) := \arg\min_{i<j\in[K]} D_{ij}(\mathcal{M})$. For each $1 \le \lambda \le w$, pick $y^*_\lambda \in \mathcal{P}_{i^*j^*}(\lambda)$ satisfying

$$\Delta_{i^*j^*}(\lambda) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i^*)) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, j^*)) \tag{7}$$

where such a y^*_λ must exist due to the following lemma whose proof is provided in Appendix C.1.

Lemma 10 *For any given $i \neq j \in [K]$ and λ, there exists $y^*_\lambda \in \mathcal{P}_{ij}(\lambda)$ such that*

$$\Delta_{ij}(\lambda) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i)) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, j)) \; .$$

Using i^*, j^* and y^*_λ, we couple the generation of ratings $\{X_{uv} : u \in \mathcal{U}, v \in \mathcal{V}\}$ under Φ and Ψ in what follows.

C1. The partition of items $\mathcal{V}_1, ..., \mathcal{V}_K$, the set of items assigned to each user $\{\mathcal{W}_u \subset \mathcal{V} : u \in \mathcal{U}\}$, and the set of items that each user gives ratings $\{\mathcal{R}_u \subset \mathcal{W}_u : u \in \mathcal{U}\}$ under Φ are the same as those generated under Ψ.

C2. We select item v^* in $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}$ uniformly at random. If $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} = \emptyset$, we select item v^* in $\mathcal{V}$ uniformly at random. The ratings by user u such that $v^* \notin \mathcal{R}_u$ generated under Ψ are the same as those generated under Φ.

C3. Let $\lambda_u := |\mathcal{R}_u|$. For each user u such that $v^* \in \mathcal{R}_u$, let $\boldsymbol{w}_u := (w_{u,1}, ..., w_{u,\lambda_u-1}, w_{u,\lambda_u} = v^*)$ be the unique sequence of items in $\mathcal{R}_u$ such that $w_{u,t} < w_{u,t+1}$ for all $t < \lambda_u - 1$ and $w_{u,\lambda_u} = v^*$. Regardless of user type, the sequence of ratings $\boldsymbol{x}_u \in \{0,1\}^{\lambda_u}$ on the sequence of items $\boldsymbol{w}_u$ generated under Ψ are observed with probability $y^*_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)$, where $\sigma(\boldsymbol{w}_u) \in [K]^{\lambda_u}$ is the sequence of clusters such that for all $t \le \lambda_u$, $\sigma(w_{u,t}) = k$ if $w_{u,t} \in \mathcal{V}_k$.

Log-likelihood ratio and its connection to the number of misclassified items. For observed ratings $\{x_{uv} : u \in \mathcal{U}, v \in \mathcal{V}\}$, we introduce the ratio of the log-likelihood of the observation under Ψ to that under Φ in the following:

$$\mathcal{G} := \sum_{u\in\mathcal{U}} \mathbb{1}_{[v^*\in\mathcal{R}_u]} \log \frac{y^*_{\lambda_u}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_{\lambda_u}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} = \sum_{u\in\mathcal{U}} \sum_{\lambda=1}^{w} \mathbb{1}_{[v^*\in\mathcal{R}_u, \lambda_u=\lambda]} \log \frac{y^*_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} \; , \tag{8}$$

where $\mathbb{1}_{\mathcal{C}}$ is the indicator of event $\mathcal{C}$ and $\boldsymbol{w}_u$, $\sigma(\boldsymbol{w}_u)$, and $\boldsymbol{x}_u$ are described in C3.

Let π be a clustering algorithm that outputs $\hat{\mathcal{V}}_1, ..., \hat{\mathcal{V}}_K$ such that $\sum_{k=1}^K \hat{\mathcal{V}}_k = \mathcal{V}$ and $\hat{\mathcal{V}}_i \cap \hat{\mathcal{V}}_j = \emptyset$ for all $i \neq j$. Without loss of generality, we assume $\left|\bigcup_{k=1}^K \hat{\mathcal{V}}_k \setminus \mathcal{V}_k\right| \le \left|\bigcup_{k=1}^K \hat{\mathcal{V}}_{\theta(k)} \setminus\right.$

$\mathcal{V}_k\big|$ for any permutation θ of $[K]$. Let $\mathcal{E}$ be the set of misclassified items by π, i.e., $\mathcal{E} = \bigcup_{k=1}^{K} \hat{\mathcal{V}}_k \setminus \mathcal{V}_k$. Then we have $|\mathcal{E}| = \varepsilon^{\pi}(n,m)$.

Let $\mathbb{P}_\Psi$ and $\mathbb{E}_\Psi$ (resp. $\mathbb{P}_\Phi$ and $\mathbb{E}_\Phi$) denote the conditional probability measure and the conditional expectation given $v^* \in \mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset$ in the perturbed model Ψ (resp. original model Φ), respectively, where we also use $\mathbb{P}$ and $\mathbb{E}$ for the probability measure and the expectation the original model Φ without the condition. We establish a connection between $\mathbb{E}[\varepsilon^{\pi}(n,m)]$ and the distribution of $\mathcal{G}$ under $\mathbb{P}_\Psi$. For any function $g(n)$, it is straightforward to check that

$$\begin{aligned}\mathbb{P}_\Psi\{\mathcal{G} \le g(n)\} &= \mathbb{P}_\Psi\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\} + \mathbb{P}_\Psi\{\mathcal{G} \le g(n), v^* \notin \mathcal{E}\} \\ &\le \underbrace{\mathbb{P}_\Psi\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\}}_{(a)} + \underbrace{\mathbb{P}_\Psi\{v^* \notin \mathcal{E}\}}_{(b)} \ .\end{aligned} \tag{9}$$

We first obtain an upper bound on (a) in (9). Using the log-likelihood ratio $\mathcal{G}$, it is not hard to check

$$\begin{aligned}\mathbb{P}_\Psi\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\} &= \int_{\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\}} d\mathbb{P}_\Psi \\ &= \int_{\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\}} \exp(\mathcal{G})\ d\mathbb{P}_\Phi \\ &\le \exp(g(n)) \cdot \mathbb{P}_\Phi\{\mathcal{G} \le g(n), v^* \in \mathcal{E}\} \\ &\le \exp(g(n)) \cdot \mathbb{P}_\Phi\{v^* \in \mathcal{E}\} \\ &\le \exp(g(n)) \cdot \frac{\mathbb{E}[\varepsilon^{\pi}(n,m)]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n} \ ,\end{aligned} \tag{10}$$

where for the last inequality, we use the fact that under the original model Φ, we cannot distinguish between v^* and any other in the same cluster which v^* belongs to. Indeed, recalling that v^* is selected in $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}$ uniformly at random if $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset$, it follows that

$$\begin{aligned}\mathbb{P}_\Phi\{v^* \in \mathcal{E}\} &= \frac{\mathbb{P}\{v^* \in \mathcal{E}, v^* \in \mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}\}}{\mathbb{P}\{v^* \in \mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset\}} \\ &= \frac{\mathbb{P}\{v \in \mathcal{E}, v \in \mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}\}}{\mathbb{P}\{\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset\}} \\ &\le \frac{\mathbb{P}\{v \in \mathcal{E}\}}{\mathbb{P}\{\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset\}} \ = \ \frac{\mathbb{E}[\varepsilon^{\pi}(n,m)]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n} \ .\end{aligned}$$

where v is selected in $\mathcal{V}$ uniformly at random.

We now obtain an upper bound on (b) in (9). Since under the perturbed model Ψ, the observed ratings do not depend on whether v^* belongs to cluster i^* or j^* if $v^* \in \mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}$, we have $\mathbb{P}_\Psi\{v^* \in \mathcal{V}^{\pi}_{i^*} \mid v^* \in \mathcal{V}_{i^*}\} = \mathbb{P}_\Psi\{v^* \in \mathcal{V}^{\pi}_{i^*} \mid v^* \in \mathcal{V}_{j^*}\}$ and

$$\mathbb{P}_\Psi\{v^* \in \mathcal{V}^{\pi}_{j^*} \mid v^* \in \mathcal{V}_{i^*}\} = \mathbb{P}_\Psi\{v^* \in \mathcal{V}^{\pi}_{j^*} \mid v^* \in \mathcal{V}_{j^*}\} \ .$$

Hence, recalling that v^* is selected in $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*}$ uniformly at random if $\mathcal{V}_{i^*} \cup \mathcal{V}_{j^*} \neq \emptyset$, it follows that

$$\mathbb{P}_\Psi\{v^* \notin \mathcal{E}\}$$

$$
\begin{aligned}
&= \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*}, v^* \in \mathcal{V}_{i^*}\} + \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{j^*}, v^* \in \mathcal{V}_{j^*}\} \\
&= \mathbb{P}_\Psi\{v^* \in \mathcal{V}_{i^*}\} \cdot \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*} \mid v^* \in \mathcal{V}_{i^*}\} + \mathbb{P}_\Psi\{v^* \in \mathcal{V}_{j^*}\} \cdot \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{j^*} \mid v^* \in \mathcal{V}_{j^*}\} \\
&= \frac{\alpha_{i^*}}{\alpha_{i^*}+\alpha_{j^*}} \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*} \mid v^* \in \mathcal{V}_{i^*}\} + \frac{\alpha_{j^*}}{\alpha_{i^*}+\alpha_{j^*}} \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{j^*} \mid v^* \in \mathcal{V}_{j^*}\} \\
&= \frac{\alpha_{i^*}}{\alpha_{i^*}+\alpha_{j^*}} \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*} \mid v^* \in \mathcal{V}_{i^*}\} + \frac{\alpha_{j^*}}{\alpha_{i^*}+\alpha_{j^*}} \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{j^*} \mid v^* \in \mathcal{V}_{i^*}\} \\
&\le \frac{\alpha_{j^*}}{\alpha_{i^*}+\alpha_{j^*}} \ , \qquad (11)
\end{aligned}
$$

where for the last inequality, we use the choice of i^*, j^* such that $i^* < j^*$, i.e., $\alpha_{i^*} \le \alpha_{j^*}$, and the fact that $\mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*} \mid v^* \in \mathcal{V}_{i^*}\} + \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{j^*} \mid v^* \in \mathcal{V}_{i^*}\} = \mathbb{P}_\Psi\{v^* \in \hat{\mathcal{V}}_{i^*} \cup \hat{\mathcal{V}}_{j^*} \mid v^* \in \mathcal{V}_{i^*}\} \le 1$.

Combining (9), (10) and (11), it follows that

$$
\begin{aligned}
\mathbb{P}_\Psi\{\mathcal{G} \le g(n)\} &\le \exp\left(g(n)\right) \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n} + \frac{\alpha_{j^*}}{\alpha_{i^*}+\alpha_{j^*}} \\
&\le \exp\left(g(n)\right) \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{(\alpha_{i^*}+\alpha_{j^*})n} + \frac{\alpha_{j^*}}{\alpha_{i^*}+\alpha_{j^*}} \ .
\end{aligned}
$$

Plugging in $g(n) = \log\left(n/\mathbb{E}[\varepsilon^\pi(n,m)]\right) - \log(2/\alpha_{i^*})$, we have

$$
\mathbb{P}_\Psi\{\mathcal{G} \le \log(n/\mathbb{E}[\varepsilon^\pi(n,m)]) - \log(2/\alpha_{i^*})\} \ \le \ 1 - \frac{\alpha_{i^*}}{2} < 1 - \frac{\alpha_{i^*}}{3} \ . \qquad (12)
$$

In addition, from Chebyshev's inequality,

$$
\mathbb{P}_\Psi\left\{\mathcal{G} \le \mathbb{E}_\Psi\left[\mathcal{G}\right] + \sqrt{\frac{3}{\alpha_i}\mathbb{E}_\Psi[(\mathcal{G}-\mathbb{E}_\Psi[\mathcal{G}])^2]}\right\} \ \ge \ 1 - \frac{\alpha_{i^*}}{3} \ . \qquad (13)
$$

From (12) and (13), it follows that

$$
\log(n/\mathbb{E}[\varepsilon^\pi(n,m)]) - \log(2/\alpha_{i^*}) \le \mathbb{E}_\Psi\left[\mathcal{G}\right] + \sqrt{\frac{3}{\alpha_i}\mathbb{E}_\Psi[(\mathcal{G}-\mathbb{E}_\Psi[\mathcal{G}])^2]}
$$

which implies that if the algorithm π satisfies $\mathbb{E}[\varepsilon^\pi(n,m)] \le s$, then

$$
\log(n/s) - \log(2/\alpha_{i^*}) \le \mathbb{E}_\Psi\left[\mathcal{G}\right] + \sqrt{\frac{3}{\alpha_i}\mathbb{E}_\Psi[(\mathcal{G}-\mathbb{E}_\Psi[\mathcal{G}])^2]} \ . \qquad (14)
$$

The log-likelihood ratio $\mathcal{G}$. To complete the proof, we now obtain the quantities of $\mathcal{G}$ in (14). Using the conditional independence of each term of the summation in (8) given the partition of items, it is not hard to check that for an arbitrary user $u \in \mathcal{U}$,

$$
\begin{aligned}
\mathbb{E}_\Psi[\mathcal{G}] &= m \cdot \sum_{\lambda=1}^{w} \mathbb{P}\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \Delta_{i^*j^*}(\lambda) \qquad (15) \\
&= m \cdot \frac{wq}{n} \cdot \sum_{\lambda=1}^{w} \binom{w-1}{\lambda-1} q^{\lambda-1}(1-q)^{w-\lambda} \cdot \Delta_{i^*j^*}(\lambda)
\end{aligned}
$$

$$= \frac{mwq}{n} \cdot \sum_{\lambda=1}^{w} B(w,\lambda,q) \cdot \Delta_{i^*j^*}(\lambda) \ , \tag{16}$$

where we provide the detailed steps for (15) in Appendix C.2. We also obtain an upper bound of the second term of (14) from Lemma 11 whose proof is provided in Appendix C.3.

Lemma 11 *Suppose p satisfies (A1). Then,*

$$\begin{aligned} &\mathbb{E}_\Psi[(\mathcal{G} - \mathbb{E}_\Psi[\mathcal{G}])^2] \\ &\le \frac{mwq}{n} \log\eta \cdot \sum_{\lambda=1}^{w} B(w,\lambda,q) \left(\Delta_{i^*j^*}(\lambda) + \sqrt{\Delta_{i^*j^*}(\lambda)}\right) \\ &\quad + m^2 \log\eta \cdot \left(\frac{wq}{n}\right)^2 \cdot \min\left\{1, \frac{((w-1)q)^2}{n-1}\right\} \cdot \left(\Delta_{i^*j^*}(w) + \sqrt{\Delta_{i^*j^*}(w)}\right) \ . \end{aligned} \tag{17}$$

Hence, applying (16) and (17) in (14), we have

$$\begin{aligned} &\log(n/s) - \log(2/\alpha_{i^*}) \\ &\le \mathbb{E}_\Psi\left[\mathcal{G}\right] + \sqrt{\frac{3}{\alpha_i}\mathbb{E}_\Psi[(\mathcal{G} - \mathbb{E}_\Psi[\mathcal{G}])^2]} \\ &\le \frac{mwq}{n} \sum_{\lambda=1}^{w} B(w,\lambda,q) \cdot \Delta_{i^*j^*}(\lambda) && \text{(18a)} \\ &\quad + \sqrt{\frac{mwq}{n}} \cdot \left(\frac{3}{\alpha_{i^*}} \log\eta \cdot \sum_{\lambda=1}^{w} B(w,\lambda,q) \left(\Delta_{i^*j^*}(\lambda) + \sqrt{\Delta_{i^*j^*}(\lambda)}\right)\right)^{1/2} && \text{(18b)} \\ &\quad + \frac{mwq}{n} \left(\frac{3}{\alpha_{i^*}} \log\eta \cdot \min\left\{1, \frac{((w-1)q)^2}{n-1}\right\} \cdot \left(\Delta_{i^*j^*}(w) + \sqrt{\Delta_{i^*j^*}(w)}\right)\right)^{1/2} \ , && \text{(18c)} \end{aligned}$$

where for the last inequality, we also use the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Since $s = o(n)$, the r.h.s. of (18) is increasing with respect to n. Hence, when $w = 1$ and $q > 0$, there exists constant $\varepsilon > 0$ such that $\Delta_{i^*j^*}(1) > \varepsilon$ due to Lemmas 8 and 9. This shows Theorem 1 when $w = 1$ since (18c) $= 0$ and (18b)/(18a) $= o(1)$.

Similarly, when $w \ge 2$, the fact that $s = o(n)$ implies that there exists constant $\varepsilon > 0$ such that $\varepsilon \le \Delta_{i^*j^*}(\lambda, p) \le \log\eta$ for all $\lambda \ge 2$. We note that

$$\sum_{\lambda=2}^{w} B(w,\lambda,q) = (1 - (1-q)^{w-1}) \ge \frac{(w-1)q}{1+(w-1)q} \ .$$

Hence, for $wq = o(\sqrt{n})$, we have

$$\text{(18a)} = \Omega\left(\frac{mwq}{n}\left(\frac{wq}{1+wq} + \Delta_{i^*j^*}(1)\right)\right) \quad \text{and} \quad \text{(18c)} = O\left(\frac{mwq}{n} \cdot \frac{wq}{\sqrt{n}}\right) \ ,$$

which implies $\liminf_{n\to\infty} \frac{mwqD(\mathcal{M})}{n\log(n/s)} \ge 1$. In addition, when $wq = \Omega(\sqrt{n})$, we have (18c) $= O\left(\frac{mwq}{n}\right)$ and ((18a)+(18b))/(18c) $= O(1)$ which implies $\log(n/s) = O\left(\frac{mwq}{n}\right)$. Hence, when $wq = \Omega(\sqrt{n})$, we have $mwq = \Omega\left(n\log(n/s)\right)$ which completes the proof of Theorem 1.

C.1. Proof of Lemma 10

We prove by contradiction that such a y^*_λ exists. Suppose that $y^*_\lambda \in \mathcal{P}_{ij}(\lambda)$ satisfies

$$\Delta_{ij}(\lambda) = \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i)) > \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, j)) \,.$$

Then there exists $\boldsymbol{k}' \in [K]^{\lambda-1}$ such that $\mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', i)\big) > \mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', j)\big)$. Since the KL divergence cannot be less than 0, this implies $y^*_\lambda(\boldsymbol{k}') \neq f_\lambda(\boldsymbol{k}', i)$. Thus, noting continuity of the KL divergence, for some $0 < \varepsilon' < \big(\mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', i)\big) - \mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', j)\big)\big)/2$, we can construct $y'_\lambda \in \mathcal{P}_{ij}(\lambda)$ such that $y'_\lambda(\boldsymbol{k}) = y^*_\lambda(\boldsymbol{k})$ for all $\boldsymbol{k} \neq \boldsymbol{k}'$,

$$\mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', i)\big) - \varepsilon' < \mathrm{KL}\big(y'_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', i)\big) < \mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', i)\big), \text{ and}$$
$$\mathrm{KL}\big(y'_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', j)\big) < \mathrm{KL}\big(y^*_\lambda(\boldsymbol{k}') \big\| f_\lambda(\boldsymbol{k}', j)\big) + \varepsilon'.$$

This construction of y'_λ implies that

$$\Delta_{ij}(\lambda) > \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}\big(y'_\lambda(\boldsymbol{k}) \big\| f_\lambda(\boldsymbol{k}, i)\big) > \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}\big(y'_\lambda(\boldsymbol{k}) \big\| f_\lambda(\boldsymbol{k}, j)\big)$$

which contradicts the definition of $\Delta_{ij}(\lambda)$ and complete the proof of Lemma 10 by contradiction.

C.2. Proof of (15)

Using the conditional independence of each term of $\mathcal{G}$ given $\tilde{\sigma} \in [K]^{\mathcal{V}}$ such that $\tilde{\sigma}(v^*) = i^*$, we have

$$\begin{aligned}
\mathbb{E}_\Psi[\mathcal{G} \mid \sigma(v^*) = i^*, \sigma = \tilde{\sigma}] &= \mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}} \sum_{\lambda=1}^{w} \mathbb{1}_{[v^*\in\mathcal{R}_u, \lambda_u = \lambda]} \log \frac{y^*_\lambda(\tilde{\sigma}(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_\lambda(\tilde{\sigma}(\boldsymbol{w}_u); \boldsymbol{x}_u)}\right] \\
&= \sum_{\lambda=1}^{w} \sum_{u\in\mathcal{U}} \mathbb{E}_\Psi\left[\mathbb{1}_{[v^*\in\mathcal{R}_u, \lambda_u = \lambda]} \log \frac{y^*_\lambda(\tilde{\sigma}(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_\lambda(\tilde{\sigma}(\boldsymbol{w}_u); \boldsymbol{x}_u)}\right] \,. \qquad (19)
\end{aligned}$$

Using the above, we can write the conditional expectation of $\mathcal{G}$ given only $\sigma(v^*) = i^*$ as

$$\begin{aligned}
&\mathbb{E}_\Psi[\mathcal{G} \mid \sigma(v^*) = i^*] \\
&= m \cdot \sum_{\lambda=1}^{w} \mathbb{E}_\Psi\left[\mathbb{1}_{[v^*\in\mathcal{R}_u, \lambda_u = \lambda]} \log \frac{y^*_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} \,\middle|\, \sigma(v^*) = i^*\right] \\
&= m \cdot \sum_{\lambda=1}^{w} \mathbb{P}_\Psi\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \mathbb{E}_\Psi\left[\log \frac{y^*_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_\lambda(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} \,\middle|\, \sigma(v^*) = i^*, v^* \in \mathcal{R}_u, \lambda_u = \lambda\right] \\
&= m \cdot \sum_{\lambda=1}^{w} \mathbb{P}\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \sum_{\boldsymbol{k}\in[K]^{\lambda-1}} \alpha_{\boldsymbol{k}} \cdot \mathrm{KL}(y^*_\lambda(\boldsymbol{k}) \| f_\lambda(\boldsymbol{k}, i^*)) \\
&= m \cdot \sum_{\lambda=1}^{w} \mathbb{P}\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \Delta_{i^*j^*}(\lambda) \,,
\end{aligned}$$

where user $u \in \mathcal{U}$ is arbitrary, and for the last equality, we use (7). Similarly to this, we can also obtain

$$\mathbb{E}_\Psi[\mathcal{G} \mid \sigma(v^*) = j^*] = m \cdot \sum_{\lambda=1}^{w} \mathbb{P}\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \Delta_{i^* j^*}(\lambda) \ ,$$

which completes the proof of (15).

C.3. Proof of Lemma 11

Let $\mathcal{G}_u := \mathbb{1}_{[v^* \in \mathcal{R}_u]} \log \frac{y^*_{\lambda_u}(\sigma'(\boldsymbol{w}_u);\boldsymbol{x}_u)}{f_{\lambda_u}(\sigma'(\boldsymbol{w}_u);\boldsymbol{x}_u)}$ so that $\mathcal{G} = \sum_{u \in \mathcal{U}} \mathcal{G}_u$. We start with the following:

$$\begin{aligned}\mathbb{E}_\Psi[\mathcal{G}^2 \mid \sigma(v^*) = i^*] &= \mathbb{E}_\Psi\left[\left(\sum_{u \in \mathcal{U}} \mathcal{G}_u\right)^2 \middle| \ \sigma(v^*) = i^*\right] = \mathbb{E}_\Psi\left[\sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}} \mathcal{G}_u \mathcal{G}_{u'} \ \middle| \ \sigma(v^*) = i^*\right] \\ &= \mathbb{E}_\Psi\left[\sum_{u \in \mathcal{U}} \mathcal{G}_u^2 \ \middle| \ \sigma(v^*) = i^*\right] + \mathbb{E}_\Psi\left[\sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}: u' \neq u} \mathcal{G}_u \mathcal{G}_{u'} \ \middle| \ \sigma(v^*) = i^*\right] \ .\end{aligned} \tag{20}$$

Recalling (6) implied from (A1), it follows that for all $\lambda \geq 1$,

$$\max_{\boldsymbol{k} \in \{0,1\}^{\lambda-1}, \boldsymbol{x} \in \{0,1\}^\lambda} \left| \log \frac{y^*_\lambda(\boldsymbol{k};\boldsymbol{x})}{f_\lambda(\boldsymbol{k};\boldsymbol{x})} \right| \leq \log \eta \ .$$

Hence, using Pinskers inequalities, it follows that

$$\begin{aligned}\mathbb{E}\left[\log^2 \frac{y^*_\lambda(\boldsymbol{k};\boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i^*;\boldsymbol{x})}\right] &\leq \log \eta \cdot \mathbb{E}\left[\left| \log \frac{y^*_\lambda(\boldsymbol{k};\boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i^*;\boldsymbol{x})} \right|\right] \\ &\leq \log \eta \cdot \left(\mathbb{E}\left[\log \frac{y^*_\lambda(\boldsymbol{k};\boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i^*;\boldsymbol{x})}\right] + \sqrt{\mathbb{E}\left[\log \frac{y^*_\lambda(\boldsymbol{k};\boldsymbol{x})}{f_\lambda(\boldsymbol{k}, i^*;\boldsymbol{x})}\right]} \right) \ ,\end{aligned} \tag{21}$$

where the expectation is taken with respect to $\boldsymbol{x} \in \{0,1\}^\lambda$ drawn from distribution $y^*_\lambda(\boldsymbol{k})$. Thus, using the similar decomposition in (19) from the conditional independence given $\sigma = \sigma'$ with (21), we have

$$\begin{aligned}\mathbb{E}_\Psi\left[\sum_{u \in \mathcal{U}} \mathcal{G}_u^2 \ \middle| \ \sigma(v^*) = i^*\right] &\leq m \log \eta \cdot \sum_{\lambda=1}^{w} \mathbb{P}\{v^* \in \mathcal{R}_u, \lambda_u = \lambda\} \cdot \left(\Delta_{i^* j^*}(\lambda) + \sqrt{\Delta_{i^* j^*}(\lambda)}\right) \\ &\leq \frac{mwq}{n} \log \eta \cdot \sum_{\lambda=1}^{w} B(w, \lambda, q) \cdot \left(\Delta_{i^* j^*}(\lambda) + \sqrt{\Delta_{i^* j^*}(\lambda)}\right)\end{aligned} \tag{22}$$

We now bound above the last term in (20). We first decompose it by separating the event when v^* is rated by both of two different users u, u' into two disjoint cases : (i) only v^* is rated by both users, i.e., $|\mathcal{R}_u \cap \mathcal{R}_{u'}| = 1$, and (ii) both users rate v^* but there are other items rated by them in common, i.e.,$|\mathcal{R}_u \cap \mathcal{R}_{u'}| \geq 2$. Formally,

$$\mathbb{E}_\Psi\left[\sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}: u' \neq u} \mathcal{G}_u \mathcal{G}_{u'} \ \middle| \ \sigma(v^*) = i^*\right]$$

$$
\begin{aligned}
&= \mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}:u'\neq u}\mathbb{1}_{[|\mathcal{R}_u\cap\mathcal{R}_{u'}|=1]}\mathcal{G}_u\mathcal{G}_{u'}\;\middle|\;\sigma(v^*)=i^*\right]\\
&\quad+\mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}:u'\neq u}\mathbb{1}_{[|\mathcal{R}_u\cap\mathcal{R}_{u'}|\geq 2]}\mathcal{G}_u\mathcal{G}_{u'}\;\middle|\;\sigma(v^*)=i^*\right]\\
&\leq\left(\sum_{u\in\mathcal{U}}\mathbb{E}_\Psi\left[\mathcal{G}_u\mid\sigma(v^*)=i^*\right]\right)^2+\mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}:u'\neq u}\mathbb{1}_{[|\mathcal{R}_u\cap\mathcal{R}_{u'}|\geq 2]}\mathcal{G}_u\mathcal{G}_{u'}\;\middle|\;\sigma(v^*)=i^*\right]\\
&=\left(\mathbb{E}_\Psi\left[\mathcal{G}\right]\right)^2+\mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}:u'\neq u}\mathbb{1}_{[|\mathcal{R}_u\cap\mathcal{R}_{u'}|\geq 2]}\mathcal{G}_u\mathcal{G}_{u'}\;\middle|\;\sigma(v^*)=i^*\right], \qquad (23)
\end{aligned}
$$

where for the last inequality, we use the fact that $\mathcal{G}_u$ and $\mathcal{G}_{u'}$ are correlated only on the cluster which v^* belongs to. Noting $\mathcal{R}_u$ is selected independently of $\mathcal{R}_{u'}$ and using (A1) and using (21), it is not hard to check that there exists a constant $C>0$ such that

$$
\begin{aligned}
&\mathbb{E}_\Psi\left[\sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}:u'\neq u}\mathbb{1}_{[|\mathcal{R}_u\cap\mathcal{R}_{u'}|\geq 2]}\mathcal{G}_u\mathcal{G}_{u'}\;\middle|\;\sigma(v^*)=i^*\right]\\
&\leq m^2\log\eta\cdot\left(\frac{wq}{n}\right)^2\cdot\min\left\{1,\frac{((w-1)q)^2}{n-1}\right\}\cdot\max_{\lambda\geq 2}\left(\Delta_{i^*j^*}(\lambda)+\sqrt{\Delta_{i^*j^*}(\lambda)}\right)\\
&\leq m^2\log\eta\cdot\left(\frac{wq}{n}\right)^2\cdot\min\left\{1,\frac{((w-1)q)^2}{n-1}\right\}\cdot\left(\Delta_{i^*j^*}(w)+\sqrt{\Delta_{i^*j^*}(w)}\right), \qquad (24)
\end{aligned}
$$

where for the first inequality, we use the fact that $\mathbb{P}\{|\mathcal{R}_u\cap\mathcal{R}_{u'}|\geq 2\mid v^*\in\mathcal{R}_u\cap\mathcal{R}_{u'}\}\leq\min\left\{1,\left(\frac{(w-1)q}{n-1}\right)^2(n-1)\right\}$ and for the last inequality, we use Lemma 8.

Combining (22), (23), and (24), we have

$$
\begin{aligned}
&\mathbb{E}_\Psi[\mathcal{G}^2\mid\sigma(v^*)=i^*]-(\mathbb{E}_\Psi[\mathcal{G}])^2\\
&\leq\frac{mwq}{n}\log\eta\cdot\sum_{\lambda=1}^{w}B(w,\lambda,q)\cdot\left(\Delta_{i^*j^*}(\lambda)+\sqrt{\Delta_{i^*j^*}(\lambda)}\right)\\
&\quad+m^2\log\eta\cdot\left(\frac{wq}{n}\right)^2\cdot\min\left\{1,\frac{((w-1)q)^2}{n-1}\right\}\cdot\left(\Delta_{i^*j^*}(w)+\sqrt{\Delta_{i^*j^*}(w)}\right),
\end{aligned}
$$

which completes the proof of Lemma 11 since we can also obtain the same upper bound for $\mathbb{E}_\Psi[\mathcal{G}^2\mid\sigma(v^*)=j^*]$ and $\mathbb{E}_\Psi[(\mathcal{G}-\mathbb{E}_\Psi[\mathcal{G}])^2]=\mathbb{E}_\Psi[\mathcal{G}^2]-(\mathbb{E}_\Psi[\mathcal{G}])^2$.

Appendix D. Proof of Theorem 2

We will use another change-of-measure argument similar to the one in the proof of Theorem 3 in (Yun and Proutiere, 2014). In the following, we refer to Φ as the true stochastic model of ratings with the random assignment. We first construct a slightly perturbed model Ψ' coupled with Φ.

Construction of Ψ'. We couple the generation of ratings under Φ and Ψ' as follows. Let $(i^*,j^*):=\arg\min_{i<j\in[K]}\tilde{D}_{ij}(\mathcal{M})$.

C1. We generate the hidden partition of items $\mathcal{V}_1, ..., \mathcal{V}_K$ under Φ. Then we select two items v_{i^*} and v_{j^*} uniformly at random from $\mathcal{V}_{i^*}$ and $\mathcal{V}_{j^*}$, respectively, if neither $\mathcal{V}_{i^*}$ nor $\mathcal{V}_{j^*}$ are empty, and we select v_{i^*} and v_{j^*} uniformly at random in $\mathcal{V}$ otherwise. The perturbed model Ψ' has the partition of items $\mathcal{V}'_1, \dots, \mathcal{V}'_K$ such that

- $\mathcal{V}'_i = \mathcal{V}_i$ for all $i \in [K] \setminus \{i^*, j^*\}$,
- $\mathcal{V}'_{i^*} = \{v_{j^*}\} \cup \mathcal{V}_{i^*} \setminus \{v_{i^*}\}$ and $\mathcal{V}'_{j^*} = \{v_{i^*}\} \cup \mathcal{V}_{j^*} \setminus \{v_{j^*}\}$.

This is a swap of two items v_{i^*} and v_{j^*}. We denote by $\sigma(v) \in [K]$ the cluster of v under Φ, i.e., $\sigma(v) = k$ if $v \in \mathcal{V}_k$. Let $\sigma'(v) \in [K]$ be the cluster of v under Ψ', i.e., $\sigma'(v) = k$ if $v \in \mathcal{V}'_k$.

C2. For each user $u \in \mathcal{U} = [m]$, the set of w items assigned to user u $\mathcal{W}_u$ and the set of items user u rate $\mathcal{R}_u$ are the same under both models Φ and Ψ', where $\mathcal{W}_u$ is selected arbitrarily (or adaptively). The ratings by user u generated under Ψ' such that $\mathcal{R}_u \cap \{v_{i^*}, v_{j^*}\} = \emptyset$ are the same as those generated under Φ.

C3. Let $\lambda_u := |\mathcal{R}_u|$. For each user u with $\mathcal{R}_u \cap \{v_{i^*}, v_{j^*}\} \neq \emptyset$, let $\boldsymbol{w}_u$ be a unique sequence of items in $\mathcal{R}_u$. Regardless of user type, the sequence of ratings $\boldsymbol{x}_u \in \{0,1\}^{\lambda_u}$ on the sequence of items $\boldsymbol{w}_u$ generated under Ψ' are observed with probability $f_{\lambda_u}(\sigma'(\boldsymbol{w}_u); \boldsymbol{x}_u)$, where $\sigma'(\boldsymbol{w}_u) \in [K]^{\lambda_u}$ is the sequence of clusters under Ψ' such that for all $t \leq \lambda_u$, $\sigma'(w_{u,t}) = k$ if $w_{u,t} \in \mathcal{V}'_k$.

Log-likelihood ratio and its connection to the number of misclassified items. For observed ratings $\{x_{uv} : u \in \mathcal{U}, v \in \mathcal{V}\}$, we introduce the ratio of the log-likelihood of the observation under Ψ' to that under Φ in the following:

$$\mathcal{G}' := \sum_{u=1}^{m} \mathbb{1}_{[\mathcal{R}_u \cap \{v_{i^*}, v_{j^*}\} \neq \emptyset]} \cdot \log \frac{f_{\lambda_u}(\sigma'(\boldsymbol{w}_u); \boldsymbol{x}_u)}{f_{\lambda_u}(\sigma(\boldsymbol{w}_u); \boldsymbol{x}_u)} \tag{25}$$

where $\sigma(\boldsymbol{w}_u) \in [K]^{\lambda_u}$ is the sequence of clusters under Φ such that for all $t \leq \lambda_u$, $\sigma(w_{u,t}) = k$ if $w_{u,t} \in \mathcal{V}_k$.

Let $\hat{\mathcal{V}}_1, ..., \hat{\mathcal{V}}_K$ be the output of the clustering algorithm π. Without loss of generality, we assume $\left|\bigcup_{k=1}^K \hat{\mathcal{V}}_k \setminus \mathcal{V}_k\right| \leq \left|\bigcup_{k=1}^K \hat{\mathcal{V}}_{\theta(k)} \setminus \mathcal{V}_k\right|$ for any permutation θ of $[K]$. Let $\mathcal{E} := \bigcup_{k=1}^K \hat{\mathcal{V}}_k \setminus \mathcal{V}_k$. Then we have $|\mathcal{E}| = \varepsilon^\pi(n,m)$. Let $\mathbb{P}_{\Psi'}$ and $\mathbb{E}_{\Psi'}$ (resp. $\mathbb{P}_\Phi$ and $\mathbb{E}_\Phi$) denote the conditional probability measure and the conditional expectation given $v_{i^*} \in \mathcal{V}_{i^*} \neq \emptyset$ and $v_{j^*} \in \mathcal{V}_{j^*} \neq \emptyset$ in the perturbed model Ψ' (resp. the original model Φ), respectively, where we also use $\mathbb{P}$ and $\mathbb{E}$ for the probability measure and the expectation the original model Φ without the condition.

Let d_v be the total number of users assigned to item v, i.e., $d_v := |\{u \in \mathcal{U} : v \in \mathcal{W}_u\}|$. Using the distribution of $d_{v_{i^*}} + d_{v_{j^*}}$, we will establish a connection of $\mathbb{E}[\varepsilon^\pi(n,m)]$ to the distribution of $\mathcal{G}'$ under $\mathbb{P}_{\Psi'}$. For any function $g(n)$ and $\zeta \geq 0$, we have

$$\begin{aligned}\mathbb{P}_\Phi\left\{d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta\right\} &= \mathbb{P}_{\Psi'}\left\{d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta\right\} \\ &= \mathbb{P}_{\Psi'}\left\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \in \mathcal{E}\right\} \qquad (26a)\end{aligned}$$

$$+\mathbb{P}_{\Psi'}\left\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \notin \mathcal{E}\right\} \tag{26b}$$

$$+\mathbb{P}_{\Psi'}\left\{\mathcal{G}' \geq g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta\right\} \tag{26c}$$

where for the first equality, we use the fact that v_{i^*} and v_{j^*} are statistically symmetric under the models Φ or Ψ'.

In what follows, we will calculate upper bounds for the terms on the r.h.s. of (26), where we plug in $g(n)$ and ζ with some functions of $\mathbb{E}[\varepsilon^\pi(n,m)]$. For the term in (26a), similarly to (10), we have

$$\begin{aligned}
\mathbb{P}_{\Psi'}\left\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \in \mathcal{E}\right\} &= \int_{\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \in \mathcal{E}\}} d\mathbb{P}_{\Psi'} \\
&= \int_{\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \in \mathcal{E}\}} \exp(\mathcal{G}') d\mathbb{P}_{\Phi} \\
&\leq \exp(g(n))\mathbb{P}_{\Phi}\left\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \in \mathcal{E}\right\} \\
&\leq \exp(g(n))\mathbb{P}_{\Phi}\left\{v_{i^*} \in \mathcal{E}\right\} \\
&\leq \exp(g(n)) \frac{\mathbb{P}\left\{v \in \mathcal{E}\right\}}{\mathbb{P}\left\{\mathcal{V}_{i^*} \neq \emptyset, \mathcal{V}_{j^*} \neq \emptyset\right\}} \\
&= \exp(g(n)) \cdot \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n} .
\end{aligned} \tag{27}$$

Noting the swap of arbitrary two items v_{i^*} and v_{j^*} in $\mathcal{V}_{i^*}$ and $\mathcal{V}_{j^*}$, i.e., $\mathbb{P}_{\Psi'}\left\{v_{i^*} \notin \hat{\mathcal{V}}_{i^*}\right\} = \mathbb{P}_{\Phi}\left\{v \notin \hat{\mathcal{V}}_{j^*}, v \in \mathcal{V}_{j^*}\right\}$ we obtain an upper bound for (26b) in the following

$$\begin{aligned}
\mathbb{P}_{\Psi'}\left\{\mathcal{G}' < g(n), d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta, v_{i^*} \notin \mathcal{E}\right\} &\leq \mathbb{P}_{\Psi'}\left\{v_{i^*} \notin \mathcal{E}\right\} = \mathbb{P}_{\Psi'}\left\{v_{i^*} \notin \hat{\mathcal{V}}_{i^*}\right\} \\
&= \mathbb{P}_{\Phi}\left\{v \notin \mathcal{V}^\pi_{j^*}, v \in \mathcal{V}_{j^*}\right\} \\
&\leq \mathbb{P}_{\Phi}\left\{v \notin \mathcal{E}\right\} = \frac{\mathbb{P}\left\{v \in \mathcal{E}\right\}}{\mathbb{P}\left\{\mathcal{V}_{i^*} \neq \emptyset, \mathcal{V}_{j^*} \neq \emptyset\right\}} \\
&= \frac{\mathbb{E}\left[\varepsilon^\pi(n,m)\right]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n} .
\end{aligned} \tag{28}$$

We present the following lemma which provides an upper bound of (26c) using Kolmogorov's inequality. The rigorous proof is provided in Appendix D.1.

Lemma 12 *For $\zeta \geq 0$, we have*

$$\mathbb{P}_{\Psi'}\left\{\mathcal{G}' \geq \frac{5}{2} q\zeta\tilde{D}(\mathcal{M})\right\} \leq \frac{8\log\eta}{\zeta q \cdot (\tilde{D}(\mathcal{M}))^2} \cdot \sum_{\lambda=1}^{w} B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda) + \sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)$$

Combining (27), (28), and Lemma 12 with $g(n) = 5/2 \cdot q\zeta\tilde{D}(\mathcal{M})$, it follows that

$$\mathbb{P}_{\Psi'}\left\{d_{v_{i^*}} + d_{v_{j^*}} \leq \zeta\right\} \leq \left(\exp(5/2 \cdot q\zeta\tilde{D}(\mathcal{M})) + 1\right) \frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)n}$$

$$+ \frac{8\log\eta}{\zeta q\cdot(\tilde{D}(\mathcal{M}))^2}\cdot\sum_{\lambda=1}^{w}B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda)+\sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)\ .$$

Plugging in $\zeta=\frac{\log\left(\frac{n}{\mathbb{E}[\varepsilon^\pi(n,m)]}\right)}{3q\tilde{D}(\mathcal{M})}$ and using the union bound, we further obtain

$$\begin{aligned}
&\mathbb{P}\left\{d_{v_{i^*}}+d_{v_{j^*}}\le\frac{\log(n/s)}{3q\tilde{D}(\mathcal{M})}\right\}\\
&\le\mathbb{P}_\Phi\left\{d_{v_{i^*}}+d_{v_{j^*}}\le\frac{\log\left(\frac{n}{\mathbb{E}[\varepsilon^\pi(n,m)]}\right)}{3q\tilde{D}(\mathcal{M})}\right\}(1-(1-(\alpha_{i^*}+\alpha_{j^*}))^n)+(1-(\alpha_{i^*}+\alpha_{j^*}))^n\\
&\le\left(\left(\frac{n}{\mathbb{E}[\varepsilon^\pi(n,m)]}\right)^{5/6}+1\right)\frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{n}+\frac{1}{1+(\alpha_{i^*}+\alpha_{j^*})n}\\
&\quad+24\log\eta\cdot\frac{\sum_{\lambda=1}^{w}B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda)+\sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)}{\tilde{D}(\mathcal{M})}\cdot\log\left(\frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{n}\right)
\end{aligned}\tag{29}$$

where the r.h.s. converges to 0 as $n\to\infty$ since $\frac{\mathbb{E}[\varepsilon^\pi(n,m)]}{n}\le\frac{s}{n}=o(1)$. This implies that

$$\mathbb{E}[d_{v_{i^*}}+d_{v_{j^*}}]\ge\frac{\log(n/s)}{3q\tilde{D}(\mathcal{M})}\ .$$

Since each user is assigned w items and the choice of v_{i^*} and v_{j^*} is uniformly at random, we also have

$$\begin{aligned}
mw=\sum_{i=1}^{K}\mathbb{E}\left[\sum_{v\in\mathcal{V}_i}d_v\right]&\ge\mathbb{E}\left[\sum_{v\in\mathcal{V}_{i^*}\cup\mathcal{V}_{j^*}}d_v\right]\\
&=\alpha_{i^*}n\cdot\mathbb{E}\left[d_{v_{i^*}}\right]+\alpha_{j^*}n\cdot\mathbb{E}\left[d_{v_{j^*}}\right]\\
&\ge\alpha_1 n\cdot\mathbb{E}\left[d_{v_{i^*}}+d_{v_{j^*}}\right]\ .
\end{aligned}$$

Consequently, a necessary condition for $\mathbb{E}[\varepsilon^\pi(n,m)]\le s$ is

$$\liminf_{n\to\infty}\frac{mwq\tilde{D}(\mathcal{M})}{(\alpha_1/3)n\log(n/s)}\ge1\ .$$

D.1. Proof of Lemma 12

Let u_τ denote the τ-th user assigned v_{i^*} or v_{j^*}, i.e., $\mathcal{W}_{u_\tau}\cap\{v_{i^*},v_{j^*}\}\neq\emptyset$ and $u_\tau<u_{\tau+1}$. Let $\mathcal{G}'(t)$ denote the sum of the first t nonzero terms in $\mathcal{G}'$ in (25). Formally,

$$\mathcal{G}'(t):=\sum_{\tau=1}^{t}\mathbb{1}_{[\mathcal{R}_{u_\tau}\cap\{v_{i^*},v_{j^*}\}\neq\emptyset]}\cdot\log\frac{f_{\lambda_{u_\tau}}(\sigma'(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}{f_{\lambda_{u_\tau}}(\sigma(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}\ .$$

From the construction of $\mathcal{G}'(t)$, we have $\mathbb{P}_{\Psi'}\left\{\mathcal{G}' \ge g(n)\right\} \le \mathbb{P}_{\Psi'}\left\{\max_{1\le t\le\zeta}\mathcal{G}'(t) \ge g(n)\right\}.$ Thus, we will focus on $\mathcal{G}'(t) = \sum_{\tau=1}^{t}\sum_{\lambda=1}^{w} h_{\tau,\lambda}$, where we define for $1 \le \tau \le t$ and $1 \le \lambda \le w$,

$$h_{\tau,\lambda} := \mathbb{1}_{\left[\mathcal{R}_{u_\tau}\cap\left\{v_{i^*},v_{j^*}\right\}\neq\emptyset,\lambda_{u_\tau}=\lambda\right]} \cdot \log\frac{f_\lambda(\sigma'(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}{f_\lambda(\sigma(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}\,.$$

Pick arbitrary $\tilde{\sigma},\tilde{\sigma}' \in [K]^{\mathcal{V}}$ such that $\tilde{\sigma}(v_{i^*}) = \tilde{\sigma}'(v_{j^*}) = i^*, \tilde{\sigma}(v_{j^*}) = \tilde{\sigma}'(v_{i^*}) = j^*$, and $\tilde{\sigma}(v) = \tilde{\sigma}'(v)$ for all $v \in \mathcal{V}\setminus\{v_{i^*},v_{j^*}\},$. Let $\mathbb{P}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}$ and $\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}$ denote the conditional measure and expectation given $\sigma = \tilde{\sigma}$ and $\sigma' = \tilde{\sigma}'$ under $\mathbb{P}_{\Psi'}$. We will bound above $\mathbb{P}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left\{\max_{1\le t\le\zeta}\mathcal{G}'(t) \ge g(n)\right\}$. To begin with, for every τ and λ, we obtain

$$\begin{aligned}\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[h_{\tau,\lambda}^2\right] - \left(\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[h_{\tau,\lambda}\right]\right)^2 &\le \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[h_{\tau,\lambda}^2\right]\\ &= q(2-q)B(w,\lambda,q)\cdot\mathbb{E}_{\Psi'}\left[\log^2\frac{f_\lambda(\tilde{\sigma}'(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}{f_\lambda(\tilde{\sigma}(\boldsymbol{w}_{u_\tau});\boldsymbol{x}_{u_\tau})}\right]\\ &\le q(2-q)B(w,\lambda,q)\log\eta\left(\tilde{\Delta}_{i^*j^*}(\lambda) + \sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right) \qquad (30)\end{aligned}$$

where for the last inequality, we use (21) and the fact that $\tilde{\Delta}_{i^*j^*}(\lambda)$ bounds above the KL divergence between $f_\lambda(\boldsymbol{k},i)$ and $f_\lambda(\boldsymbol{k},j)$ for every $\boldsymbol{k} \in [K]^{\lambda-1}$. Since for given τ, only one of $h_{\tau,\lambda}$'s can be non-zero, the upper bound in (30) implies

$$\begin{aligned}&\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\left(\sum_{\lambda=1}^{w}h_{\tau,\lambda} - \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\sum_{\lambda=1}^{w}h_{\tau,\lambda}\right]\right)^2\right]\\ &\le \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\left(\sum_{\lambda=1}^{w}h_{\tau,\lambda}\right)^2\right] = \sum_{\lambda=1}^{w}\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[h_{\tau,\lambda}^2\right]\\ &\le q(2-q)\log\eta\sum_{\lambda=1}^{w}B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda) + \sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)\,, \qquad (31)\end{aligned}$$

which is bounded by a constant since $\tilde{\Delta}_{i^*j^*}(\lambda) \le \log\eta < \infty$. Note that $\left(\sum_{\lambda=1}^{w}h_{\tau,\lambda}\right)_{\tau=1,..,t}$ are conditionally independent to each other given $\sigma = \tilde{\sigma}$ and $\sigma' = \tilde{\sigma}'$. Thus, from Kolmogorov's inequality, it follows that for $\gamma > 0$,

$$\begin{aligned}&\mathbb{P}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left\{\max_{1\le t\le\zeta}\left|\mathcal{G}'(t) - \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\mathcal{G}'(t)\right]\right| \ge \gamma\right\}\\ &\le \frac{\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\left(\mathcal{G}'(\zeta) - \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\mathcal{G}\right]'(\zeta)\right)^2\right]}{\gamma^2}\\ &= \frac{1}{\gamma^2}\cdot\sum_{\tau=1}^{\zeta}\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\left(\sum_{\lambda=1}^{w}h_{\tau,\lambda} - \mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\sum_{\lambda=1}^{w}h_{\tau,\lambda}\right]\right)^2\right]\,. \qquad (32)\end{aligned}$$

Since for all $t \le \zeta$, $\mathbb{E}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left[\mathcal{G}'(t)\right] \le \zeta\cdot q(2-q)\tilde{D}(\mathcal{M})$, combining (31) and (32) and plugging in $\gamma = \zeta q(1/2+q)\tilde{D}(\mathcal{M})$, we have

$$\mathbb{P}_{\Psi'|\tilde{\sigma},\tilde{\sigma}'}\left\{\max_{1\le t\le\zeta}\mathcal{G}'(t) \ge \zeta\cdot\frac{5}{2}q\tilde{D}(\mathcal{M})\right\}$$

$$\le \frac{\log\eta(2-q)\cdot\sum_{\lambda=1}^{w}B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda)+\sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)}{\zeta q\cdot((1/2+q)\tilde{D}(\mathcal{M}))^2}$$

$$\le \frac{8\log\eta\cdot\sum_{\lambda=1}^{w}B(w,\lambda,q)\left(\tilde{\Delta}_{i^*j^*}(\lambda)+\sqrt{\tilde{\Delta}_{i^*j^*}(\lambda)}\right)}{\zeta q\cdot(\tilde{D}(\mathcal{M}))^2}\,, \tag{33}$$

which completes the proof of Lemma 12 since from the construction of $\mathcal{G}'(t)$, we have $\mathbb{P}_{\Psi'}\{\mathcal{G}'\ge g(n)\}\le\mathbb{P}_{\Psi'}\{\max_{1\le t\le\zeta}\mathcal{G}'(t)\ge g(n)\}$.

Appendix E. Spectral Decomposition Algorithm

Algorithm 3 Spectral decomposition

```
Input: $A_\Gamma, \tilde{p}$
$\hat{A} \leftarrow K$-rank approximation of $A_\Gamma$
for $t = 1$ to $\log n$ do
    $Q_{t,v} \leftarrow \{w \in V : \|\hat{A}_w - \hat{A}_v\|^2 \le t\frac{\tilde{p}}{100}\}$
    $T_{t,0} \leftarrow \emptyset$
    for $k = 1$ to $K$ do
        $v_k^* \leftarrow \arg\max_v |Q_{t,v} \setminus \bigcup_{i=1}^{k-1} T_{t,i}|$
        $T_{t,k} \leftarrow Q_{t,v_k^*} \setminus \bigcup_{i=1}^{k-1} T_{t,i}$ and $\xi_{t,k} \leftarrow \sum_{v\in T_{t,k}} \hat{A}_v/|T_{t,k}|$.
    end for
    for $v \in V \setminus (\bigcup_{k=1}^{K} T_{t,k})$ do
        $k^* \leftarrow \arg\min_k \|\hat{A}_v - \xi_{t,k}\|$
        $T_{t,k^*} \leftarrow T_{t,k^*} \cup \{v\}$
    end for
    $r_t \leftarrow \sum_{k=1}^{K}\sum_{v\in T_{t,k}} \|\hat{A}_v - \xi_{t,k}\|^2$
end for
$t^* \leftarrow \arg\min_t r_t$.
$\mathcal{S}_k \leftarrow T_{t^*,k}$ for all $k$
Output: $(\mathcal{S}_k)_{k=1,\dots,K}$.
```

Appendix F. Preliminary to Spectral Analysis (Appendices G and H)

In this section, we will provide useful notations and properties for Theorems 6 and 7 (in Appendices G and H, resp.). In particular, we analyze the output A of Algorithm 1, where the analysis is analog to the one in (Yun and Proutiere, 2014, 2016). However, our analysis has to handle additional challenges mainly from dependency among entries of A. In our model, a user's ratings can be counted in several entries of A. Thus, the entries of A are dependent in a sense that the w items assigned to the user must be different to each other. To overcome this challenge, we approximate the original generation of A using two slightly perturbed generations of A, denoted by Ψ_1 and Ψ_2, where the entries of A are independent to each other. We will describe Ψ_1 and Ψ_2 further.

Notation. We first introduce the notations used in the following sections including Appendices G and H. For the sake of simplicity, we omit the conditions assumed in Theorems 6 and 7, i.e., $D(\mathcal{M}) > 0$, $h = o(\sqrt{n})$ and $h^2 m = o(n^2)$ (or equivalently $wq = o(\sqrt{n})$ and $(wq)^2 m = o(n^2)$), from the statements of upcoming lemmas. In addition, we further fix the configuration σ of clusters $\mathcal{V}_1, ..., \mathcal{V}_K$ such that $\bigcup_{k\in[K]} \mathcal{V}_k = \mathcal{V}$, $|\mathcal{V}_k| = \alpha_k n$ for all k, and if $\sigma(v) = k$, then $v \in \mathcal{V}_k$. Indeed, using the Chernoff bound, it is easy to check that for all $k \in [K]$, $||\mathcal{V}_k| - \alpha_k n| = o(n)$ with high probability since α_k is constant. Hence, we denote by $\mathbb{P}$ and $\mathbb{E}$ the (conditional) probability measure and expectation, respectively, associated with the original generation (Algorithm 1) given the fixed configuration σ.

Let $M := \mathbb{E}[A]$ and $M_\Gamma := \mathbb{E}[A_\Gamma]$, where Γ is the subset of items after trimming. We denote by A_v and M_v (resp. $A_{\Gamma,v}$ and $M_{\Gamma,v}$) the item v's column of A and M (resp. A_Γ and M_Γ), respectively. Define $p_{\max} := \max_{(v,v')\in\mathcal{V}\times\mathcal{V}} M_{vv'}$. We now define the perturbed generations of A under Ψ_1 and Ψ_2, where each entry in A is generated independently:

Generation of A by Ψ_1. There are m users each of which has h attempts to add weights to A. Starting with $A = 0$, at each attempt of each user, we add one to $A_{vv'}$ with probability $\frac{M_{vv'}}{mh}$ for all $v, v' \in \mathcal{V}$.

Generation of A by Ψ_2. There are m users each of which has h attempts to add weights to A. Starting with $A = 0$, at each attempt of each user, we add one to $A_{vv'}$ with probability $\frac{M_{vv'}}{mh}\left(1 + \frac{h}{\sqrt{n}}\right)$ for all $v, v' \in \mathcal{V}$.

Let $\mathbb{P}_1$ and $\mathbb{E}_1$ (resp. $\mathbb{P}_2$ and $\mathbb{E}_2$) denote the probability and expectation, respectively, under Ψ_1 (resp. Ψ_2). We establish the following connections of the perturbed $\mathbb{P}_1$ and $\mathbb{P}_2$ to the original $\mathbb{P}$.

Lemma 13 *For every $\mathcal{A} \subset \mathbb{Z}^{n\times n}$,*

$$\mathbb{P}\{A \in \mathcal{A}\} \leq \exp\left(O\left(\frac{h^2 m}{n}\right)\right) \mathbb{P}_1\{A \in \mathcal{A}\} .$$

Lemma 14 *For any given $c_{vv'} \geq 0$ for all $v, v' \in \mathcal{V}$ and $C \geq 0$,*

$$\mathbb{P}\left\{\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} c_{vv'} A_{vv'} \geq C\right\} \leq \mathbb{P}_2\left\{\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} c_{vv'} A_{vv'} \geq C\right\} .$$

The proofs of Lemmas 13 and 14 are provided in Appendices F.1 and F.2.

F.1. Proof of Lemma 13

Let $A(u)$ denote the weight matrix of a set of edges generated by user u in Algorithm 1, where each user attempts h times to add weights. Then,

$$A = \sum_{u=1}^{m} A(u) .$$

The number of entries that user u is able to add at the t-th attempt is $(n-2t)(n-2t-1)/2$ in Algorithm 1. However, this number under Ψ_1 is always $n(n-1)/2$. Using this, it is not hard to check that for all feasible $a \in \mathbb{Z}^{n\times n}$,

$$\begin{aligned}\mathbb{P}\left\{A(t)=a\right\} &\leq \left(1-O\left(\frac{h}{n}\right)\right)^{-h} \mathbb{P}_1\left\{A(t)=a\right\} &(34)\\ &\leq \exp\left(O\left(\frac{h^2}{n}\right)\right)\mathbb{P}_1\left\{A(t)=a\right\}, &(35)\end{aligned}$$

where we use the assumption on $h = o(\sqrt{n})$ for the last inequality. Therefore,

$$\begin{aligned}\mathbb{P}\left\{A\in\mathcal{A}\right\} &= \sum_{a_1,\dots,a_m:\sum_{u=1}^m a_u\in\mathcal{A}} \prod_{t=1}^m \mathbb{P}\left\{A(t)=a_u\right\}\\ &\leq \exp\left(O\left(\frac{mh^2}{n}\right)\right) \sum_{a_1,\dots,a_m:\sum_{u=1}^m a_u\in\mathcal{A}} \prod_{t=1}^m \mathbb{P}_1\left\{A(t)=a_u\right\}\\ &\leq \exp\left(O\left(\frac{mh^2}{n}\right)\right)\mathbb{P}_1\left\{A\in\mathcal{A}\right\},\end{aligned}$$

which completes the proof of Lemma 13.

F.2. Proof of Lemma 14

Let $A(u,t)$ be the weight matrix added by the t-th partition of the u-th user. Recalling the same logic used for (34), we also have

$$\begin{aligned}\mathbb{P}\left\{A(u,t)|A(u,t-1),\dots,A(u,1)\right\} &\leq \left(1+\frac{h}{\sqrt{n}}\right)\mathbb{P}_1\left\{A(u,t)|A(u,t-1),\dots,A(u,1)\right\}\\ &= \mathbb{P}_2\left\{A(u,t)|A(u,t-1),\dots,A(u,1)\right\}\ , &(36)\end{aligned}$$

where for the last equality, we use the fact that the probability of incrementing an entry of A at each attempt under Ψ_2 is $\left(1+\frac{h}{\sqrt{n}}\right)$-times higher than that under Ψ_1. Using (36) and noting that every entry of A is non-negative, it is straightforward to show the stochastic dominance of the weighted summation of A under Ψ_2 over that under the original model, i.e.,

$$\mathbb{P}\left\{\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} c_{vv'}A_{vv'} \geq C\right\} \leq \mathbb{P}_2\left\{\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} c_{vv'}A_{vv'} \geq C\right\}\ ,$$

which completes the proof of Lemma 14.

Appendix G. Proof of Theorem 6

In Algorithm 2, we first run Algorithm 1 to obtain weight matrix A, where we estimate the similarity between two items v, v' at $A_{vv'}$. Then, we obtain $\hat{A}$, which is the K-rank approximation of the trimmed A_Γ, and run Algorithm 3 extracting the hidden clusters based on the distance between columns of $\hat{A}$. Hence, we will use some properties of the column distance to conclude Theorem 6. (We refer to Appendix F for the definitions of notations used in the followings.)

We first study the column distance in the average matrix M_Γ. When two items x, y belong to the same cluster, it is clear that $M_{\Gamma,x} = M_{\Gamma,y}$, and when they belong to different clusters, the column distance $\|M_{\Gamma,x} - M_{\Gamma,y}\|_2^2$ might be large. More formally, we obtain the following lemma, whose proof is provided in Appendix G.1.

Lemma 15 *For all $x, y \in \Gamma$ such that $\sigma(x) \neq \sigma(y)$,*

$$\|M_{\Gamma,x} - M_{\Gamma,y}\|_2^2 = \Omega\left(n\left(\frac{wq(1 \wedge wq)m}{n^2}\right)^2\right) . \tag{37}$$

Furthermore, if item $v \in \Gamma$ is misclassified, then we should have:

$$\|\hat{A}_v - M_{\Gamma,v}\|_2^2 = \Omega\left(n\left(\frac{wq(1 \wedge wq)m}{n^2}\right)^2\right) . \tag{38}$$

To complete the proof of Theorem 6, we will obtain an upper bound on the total column distance between $\hat{A}$ and M_Γ. Note that the ranks of both $\hat{A}$ and M_Γ are K. Hence, the rank of $\hat{A} - M_\Gamma$ is less than $2K$. Using the property of Frobenius norm: if the rank of matrix A is K, $\|A\|_F^2 \leq K\|A\|_2^2$, it follows that

$$\begin{aligned}\sum_{v\in\Gamma} \|\hat{A}_v - M_{\Gamma,v}\|_2^2 &= \|\hat{A} - M_\Gamma\|_F^2 \\ &\leq 2K\|\hat{A} - M_\Gamma\|_2^2 \\ &\leq 2K\|\hat{A} - A_\Gamma\|_2^2 + 2K\|A_\Gamma - M_\Gamma\|_2^2 \\ &\leq 4K\|A_\Gamma - M_\Gamma\|_2^2 ,\end{aligned} \tag{39}$$

where we use the property of Frobenius norm for the first inequality; the triangle inequality for the second one; and the definition of $\hat{A}$, i.e., $\|\hat{A} - A_\Gamma\|_2 = \min_{M':\text{rank}(M')\leq K} \|M' - A_\Gamma\|_2 \leq \|M_\Gamma - A_\Gamma\|_2$ for the last one.

We now obtain an upper bound on the last term in (39) in the following lemma, whose proof is provided in Appendix G.2.

Lemma 16 *With high probability,*

$$\|A_\Gamma - M_\Gamma\|_2 = O\left(\sqrt{\frac{wmq(1 \wedge wq)}{n}}\right) .$$

From Lemma 16 and (39), we obtain an upper bound on the total column distance between $\hat{A}$ and M_Γ, i.e., with high probability,

$$\|\hat{A} - M_\Gamma\|_F^2 = \sum_{v\in\Gamma} \|\hat{A}_v - M_{\Gamma,v}\|_2^2 = O\left(\frac{n^2}{wq(1 \wedge wp)m}\right) . \tag{40}$$

Finally, combining (38) and (40), we bound the number of misclassified items by Algorithm 3 as follows:

$$\sum_{k=1}^{K} |\mathcal{V}_k \setminus \mathcal{S}_k| = O\left(\frac{\frac{n^2}{wq(1\wedge wp)m}}{n\left(\frac{wq(1\wedge wq)m}{n^2}\right)^2} \right)$$
$$= O\left(\frac{n^2}{wq(1\wedge wp)m} \right) ,$$

which completes the proof of Theorem 6.

G.1. Proof of Lemma 15

We will focus on the proof of (37) since from (37), it is not hard to derive (38) through the same logic used for Theorem 6 in (Yun and Proutiere, 2016). To begin with, we first bound the loss of weights after the trimming in the following lemma, whose proof is provided in Appendix G.3.

Lemma 17 *For any trimmed set* Γ*,*

$$\mathbb{P}\left\{ \sum_{v'\in\mathcal{V}\setminus\Gamma} \sum_{v\in\mathcal{V}} A_{vv'} \geq n \right\} \leq \exp(-n) .$$

Hence, it is enough to bound the column distance in M instead of M_Γ. We will show

$$\|M_x - M_y\|_2^2 \geq Cn \left(\frac{wq(1\wedge wq)m}{n^2} \right)^2 . \tag{41}$$

We denote by ρ_{ij} the expected weight between an item pair v, v' such that $v \in \mathcal{V}_i$ and $v' \in \mathcal{V}_j$, i.e., $\rho_{ij} = M_{vv'}$. Then,

$$\rho_{ij} = mh\binom{n-2}{\gamma-2}\binom{n}{\gamma}^{-1} q^2 \sum_{\ell=1}^{L} \beta_\ell p_{\ell i} p_{\ell j} Q_\ell (1+o(1))$$
$$= mh\frac{\gamma(\gamma-1)}{n(n-1)} q^2 \sum_{\ell=1}^{L} \beta_\ell p_{\ell i} p_{\ell j} Q_\ell (1+o(1)) , \tag{42}$$

where Q_ℓ is the probability that randomly selected $\gamma - 2$ items do not have any positive rating by type-ℓ user. Note that Q_ℓ is a constant for all ℓ since we set γ so that the expected number of positive ratings of each group becomes less than 2. From (42), it follows that

$$\|M_x - M_y\|_2^2 = \sum_{k=1}^{K} \sum_{v\in\mathcal{V}_k} (M_{xv} - M_{yv})^2$$
$$\geq \sum_{v\in\mathcal{V}_i} (M_{xv} - M_{yv})^2 + \sum_{v\in\mathcal{V}_j} (M_{xv} - M_{yv})^2$$
$$\geq (|\mathcal{V}_i| \wedge |\mathcal{V}_j|) \left((\rho_{ii} - \rho_{ij})^2 + (\rho_{ij} - \rho_{jj})^2 \right)$$

$$
\begin{aligned}
&\geq \frac{(|\mathcal{V}_i| \wedge |\mathcal{V}_j|)}{4} \left(\rho_{ii} - 2\rho_{ij} - \rho_{jj}\right)^2 \\
&= \frac{(|\mathcal{V}_i| \wedge |\mathcal{V}_j|)}{4} h^2 \frac{\gamma^2(\gamma-1)^2}{n^2(n-1)^2} m^2 q^4 \left(\sum_{\ell=1}^{L} \beta_\ell (p_{\ell i} - p_{\ell j})^2 Q_\ell (1 + o(1)) \right)^2 .
\end{aligned}
$$

Thus, there exists a constant $C_1 > 0$ such that

$$
\|M_x - M_y\|_2^2 \geq C_1 n \left(\frac{wm\gamma q^2}{n^2} \right)^2 ,
$$

which completes the proof of Lemma 15 since from the definition of γ, there exists a constant $C_2 > 0$ such that

$$
\gamma q \geq C_2 (1 \wedge wq) .
$$

G.2. Proof of Lemma 16

We can prove this lemma using the strategy in (Feige and Ofek, 2005). To this end, we will prove that for all $x, y \in \mathbb{R}^n$ such that $\|x\|_2 = \|y\|_2 = 1$,

$$
\left| x^\top (A - M) y \right| = O\left(\sqrt{\frac{wmq(1 \wedge wq)}{n}} \right) . \tag{43}
$$

Let $\overline{\Gamma} := \mathcal{V} \setminus \Gamma$. For any given $x, y \in \mathbb{R}^n$ such that $\|x\|_2 = \|y\|_2 = 1$, define

$$
\mathcal{L} = \left\{ (v, v') \in \mathcal{V} \times \mathcal{V} : |x_v y_{v'}| \leq \sqrt{\frac{p_{\max}}{n}} \right\} \text{ and } \overline{\mathcal{L}} = \left\{ (v, v') \in \mathcal{V} \times \mathcal{V} : |x_v y_{v'}| > \sqrt{\frac{p_{\max}}{n}} \right\} .
$$

Then, using the definitions of $\mathcal{L}$ and $\overline{\mathcal{L}}$ and the triangle inequality, we bound $\left| x^\top (A - M) y \right|$ as follows:

$$
\begin{aligned}
\left| x^\top (A - M) y \right| &\leq 2 \left| \sum_{(v,v') \in (\overline{\Gamma} \times \mathcal{V}) \cap \mathcal{L}} x_v^\top A_{vv'} y_{v'} \right| + \left| \sum_{(v,v') \in \mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \right| + \left| \sum_{(v,v') \in \overline{\mathcal{L}}} x_v A_{vv'} y_{v'} \right| \\
&\leq 2 \sqrt{\frac{p_{\max}}{n}} \left| \sum_{(v,v') \in (\overline{\Gamma} \times \mathcal{V}) \cap \mathcal{L}} A_{vv'} \right| + \left| \sum_{(v,v') \in \mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \right| + \left| \sum_{(v,v') \in \overline{L}} x_v A_{vv'} y_{v'} \right| .
\end{aligned}
$$

To prove (43), we will show that for $x, y \in \mathbb{R}^n$ such that $\|x\|_2 = \|y\|_2 = 1$,

$$
\sqrt{\frac{p_{\max}}{n}} \sum_{(v,v') \in (\overline{\Gamma} \times \mathcal{V}) \cap \mathcal{L}} A_{vv'} = O\left(\sqrt{n p_{\max}} \right) , \tag{44}
$$

$$
\sum_{(v,v') \in \mathcal{V}} x_v A_{vv'} y_{v'} - x^\top M y = O\left(\sqrt{n p_{\max}} \right) , \tag{45}
$$

$$
\sum_{(v,v') \in \overline{\mathcal{L}}} x_v A_{vv'} y_{v'} = O\left(\sqrt{n p_{\max}} \right) . \tag{46}
$$

We note that (44) is a direct consequences of Lemma 17. Using Lemma 13 of (Yun and Proutiere, 2014), it is straightforward to check (46) under $\mathbb{P}_2$. Hence, due to Lemma 14 connecting $\mathbb{P}$ and $\mathbb{P}_2$, we also have (46) under the original generation of A. Therefore, we focus on the proof of (45) to complete the proof of Lemma 16.

For given $x, y \in \mathbb{R}^{\mathcal{V}}$ such that $\|x\|_2 = \|y\|_2 = 1$, with $\lambda = \frac{1}{2}\sqrt{\frac{n}{p_{\max}}}$, we have

$$\begin{aligned}
\mathbb{P}_1\left\{\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \geq C\sqrt{np_{\max}}\right\} &\leq \frac{\mathbb{E}_1\left[\exp\left(\lambda \sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - \lambda x^\top M y\right)\right]}{\exp\left(\lambda C\sqrt{np_{\max}}\right)} \\
&= \frac{\left(1 + \sum_{(v,v')\in\mathcal{L}} \frac{M_{vv'}}{mh}\left(\exp\left(\lambda x_v y_{v'}\right) - 1\right)\right)^{mh}}{\exp\left(\lambda C\sqrt{np_{\max}} + \lambda x^\top M y\right)} \\
&\leq \frac{\left(1 + \sum_{(v,v')\in L} \frac{M-vv'}{mh}\left(\lambda x_v y_{v'} + 2(\lambda x_v y_{v'})^2\right)\right)^{mh}}{\exp\left(\lambda C\sqrt{np_{\max}} + \lambda x^\top M y\right)} \qquad (47) \\
&\leq \frac{\exp\left(\sum_{(v,v')\in L} M_{vv'}\left(\lambda x_v y_{v'} + 2(\lambda x_v y_{v'})^2\right)\right)}{\exp\left(\lambda C\sqrt{np_{\max}} + \lambda x^\top M y\right)}, \qquad (48)
\end{aligned}$$

where we use $e^x \leq x + 2x^2$ for $|x| \leq 1/2$ and $1 + x \leq e^x$ for (47) and (48), respectively.

We now bound the last term. Note that $\sum_{(v,v')\in\overline{\mathcal{L}}} |x_v y_{v'}| \leq \sqrt{\frac{n}{p_{\max}}}$ since $\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} x_v^2 y_{v'}^2 = 1$ and $|x_v y_{v'}| > \sqrt{p_{\max}/n}$ for all $(v, v') \in \overline{\mathcal{L}}$, . Thus, we have

$$\begin{aligned}
\exp\left(\sum_{(v,v')\in\mathcal{L}} \lambda M_{vv'} x_v y_{v'} - \lambda x^\top M y\right) &= \exp\left(-\sum_{(v,v)\in\overline{\mathcal{L}}} \lambda M_{vv'} x_v y_{v'}\right) \\
&\leq \exp\left(\lambda p_{\max}\sqrt{\frac{n}{p_{\max}}}\right) \\
&= \exp\left(\frac{n}{2}\right). \qquad (49)
\end{aligned}$$

In addition, recalling $\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} x_v^2 y_{v'}^2 = 1$, it also follows that

$$\exp\left(\sum_{(v,v')\in\mathcal{L}} M_{vv'} 2(\lambda x_v y_{v'})^2\right) \leq \exp\left(\frac{n}{2}\right). \qquad (50)$$

Combining (49) and (50) to (48), we have

$$\mathbb{P}_1\left\{\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \geq C\sqrt{np_{\max}}\right\} \leq \exp\left(n - \frac{C}{2}n\right).$$

Using Lemma 13, we further replace $\mathbb{P}_1$ with $\mathbb{P}$ as follows:

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \geq C\sqrt{np_{\max}}\right\}\\
&\leq \exp\left(O\left(\frac{h^2 m}{n}\right)\right)\mathbb{P}_1\left\{\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \geq C\sqrt{np_{\max}}\right\}\\
&\leq \exp\left(n + O\left(\frac{h^2 m}{n}\right) - \frac{C}{2}n\right) . \qquad (51)
\end{aligned}$$

Although $x, y \in \mathbb{R}^n$ are in a continuous space, we can use the union bound with a discrete space T_ε defined as

$$T_\varepsilon := \left\{x \in \left(\frac{\varepsilon}{\sqrt{n}}\mathbb{Z}\right)^n : \|x\|_2 \leq 1\right\} \quad \text{where} \quad 0 < \varepsilon < 1 .$$

Let $T = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. From Claim 2.4 and Claim 2.9 of (Feige and Ofek, 2005), it follows that there is a constant c_ε such that $|T_\varepsilon| \leq e^{c_\varepsilon n}$ and

$$\max_{x,y\in T} x^\top A y \leq \frac{1}{(1-\varepsilon)^2}\max_{x,y\in T_\varepsilon} x^\top A y .$$

Therefore, using the union bound and (51), we have that with probability at least $1 - \exp\left(c_\varepsilon n + n + 5\frac{h^2 m}{n} - \frac{C}{2}n\right)$,

$$\max_{x,y\in T}\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y \leq 4\max_{x,y\in T_\varepsilon}\sum_{(v,v')\in\mathcal{L}} x_v A_{vv'} y_{v'} - x^\top M y = O\left(\sqrt{np_{\max}}\right) ,$$

which implies (45) by selecting sufficiently large constant C. This completes the proof of Lemma 16.

G.3. Proof of Lemma 17

For any given set $\overline{\Gamma} \subset \mathcal{V}$ such that $|\overline{\Gamma}| = \lfloor n\exp(-n\tilde{p})\rfloor$, from Lemma 14,

$$\begin{aligned}
\mathbb{P}\left\{\sum_{v\in\overline{\Gamma}}\sum_{v'\in\mathcal{V}} A_{vv'} \geq n\right\} &\leq \mathbb{P}_2\left\{\sum_{v\in\overline{\Gamma}}\sum_{v'\in\mathcal{V}} A_{vv'} \geq n\right\}\\
&\leq \max_{\lambda>0}\frac{\mathbb{E}_2\left[\exp\left(\lambda\sum_{v\in\overline{\Gamma}}\sum_{v'\in\mathcal{V}} A_{vv'}\right)\right]}{-\lambda n}\\
&\leq \max_{\lambda>0}\frac{\exp\left(2n\lfloor n\exp(-n\tilde{p})\rfloor p_{\max}(e^\lambda - 1)\right)}{\exp(\lambda n)}\\
&\leq \exp\left(2n\lfloor n\exp(-n\tilde{p})\rfloor p_{\max}(e^5 - 1) - 5n\right)\\
&\leq \exp(-4n) , \qquad (52)
\end{aligned}$$

where the last inequality stems from the fact that $np_{\max}\exp\left(-n\tilde{p}\right) = o(1)$ when $np_{\max} = \omega(1)$. Noting that there are 2^n subsets of $\mathcal{V}$ and using the union bound and (52), it follows that

$$\mathbb{P}\left\{\sum_{v\in\mathcal{V}\setminus\Gamma}\sum_{v'\in\mathcal{V}} A_{vv'} \geq n\right\} \leq \exp(-n)\ ,$$

which completes the proof of Lemma 17.

Appendix H. Proof of Theorem 7

We will first define a subset $\mathcal{H}$ of $\mathcal{V}$ satisfying a set of criteria and show that all items in $\mathcal{H}$ are accurately classified after $\log n$ iterations of the improvement step. Then, we can complete the proof of Theorem 7 by obtaining an upper bound of $|\mathcal{V}\setminus\mathcal{H}|$ since the total number of misclassified items is less than $|\mathcal{V}\setminus\mathcal{H}|$.

For $\mathcal{A},\mathcal{B}\subset\mathcal{V}$, let $e(\mathcal{A},\mathcal{B})$ denote the summation of weights on every edge between $\mathcal{A}$ and $\mathcal{B}$, i.e., $e(\mathcal{A},\mathcal{B}) := \sum_{v\in\mathcal{A}}\sum_{v'\in\mathcal{B}} A_{vv'}$. Define $\mathcal{H}$ as the largest set of item v satisfying (H1)-(H3) in the followings:

(H1) $e(v,\mathcal{V}) \leq 10np_{\max}$.

(H2) If $v\in\mathcal{V}_i$, then

$$\sum_{k=0}^{K} e(\{v\},\mathcal{V}_k)\log\frac{\hat{p}(i,k)}{\hat{p}(j,k)} \geq \frac{np_{\max}}{\log(np_{\max})} \quad \text{for all} \quad i\neq j\ ,$$

where with a slight abuse of notation, we denote $e(\{v\},\mathcal{V}_0) := m - \sum_{k=1}^{K} e(\{v\},\mathcal{V}_k)$ and $\hat{p}(i,0) := 1-\sum_{k=1}^{K}\hat{p}(i,k)$.

(H3) $e(\{v\},\mathcal{V}\setminus\mathcal{H}) \leq 2\log^2(np_{\max})$.

Now, we will show that all items in $\mathcal{H}$ are classified correctly after $\log n$ iterations. Let $\mathcal{E}_{ij}^{(t)} := (\mathcal{S}_i^{(t)}\cap\mathcal{V}_j)\cap\mathcal{H}$ and $\mathcal{E}^{(t)} := \bigcup_{i,j:i\neq j}\mathcal{E}_{ij}^{(t)}$. At every improvement iteration, we move each item to cluster k^* having the highest (estimated) log-likelihood defined in Algorithm 2. Thus, for all t,

$$\begin{aligned}
0 &\leq \sum_{i,j:i\neq j}\sum_{v\in\mathcal{E}_{ij}^{(t)}}\sum_{k=0}^{K} e(\{v\},S_k^{(t-1)})\log\frac{\hat{p}(i,k)}{\hat{p}(j,k)}\\
&\overset{(a)}{=} \sum_{i,j:i\neq j}\sum_{v\in\mathcal{E}_{ij}^{(t)}}\sum_{k=0}^{K} e(\{v\},\mathcal{V}_k)\log\frac{\hat{p}(i,k)}{\hat{p}(j,k)} + O\left(e(\mathcal{E}^{(t)},\mathcal{E}^{(t-1)}) + e(\mathcal{E}^{(t)},\mathcal{V}\setminus\mathcal{H})\right)\\
&\overset{(b)}{\leq} -\frac{np_{\max}}{\log(np_{\max})}|\mathcal{E}^{(t)}| + O\left(e(\mathcal{E}^{(t)},\mathcal{E}^{(t-1)}) + e(\mathcal{E}^{(t)},\mathcal{V}\setminus\mathcal{H})\right)\\
&\overset{(c)}{=} -\frac{np_{\max}}{\log(np_{\max})}|\mathcal{E}^{(t)}| + O\left(\sqrt{|\mathcal{E}^{(t)}||\mathcal{E}^{(t-1)}|np_{\max}} + |\mathcal{E}^{(t)}|\log^2(np_{\max})\right)\ .
\end{aligned} \tag{53}$$

where (a) stems from the fact that $\frac{\hat{p}(j,i)}{\hat{p}(k,i)} = O(1)$ for all $i, j, k \in [K]$; (b) is obtained from (H2) and $v \in \mathcal{E}^{(t+1)}$; and (c) holds because of (H3) and Lemma 16. For the sake of simplicity, we omit the detail steps in the above inequalities since they are analog to those used for the proof of Theorem 2 in (Yun and Proutiere, 2016).

From (53), it follows that

$$\frac{|\mathcal{E}^{(t)}|}{|\mathcal{E}^{(t-1)}|} = O\left(\frac{\log(np_{\max})^2}{np_{\max}}\right) ,$$

which implies that after $\log n$ iterations, $|\mathcal{E}^{(\log n)}|$ has to be 0, i.e., items in $\mathcal{H}$ are correctly classified.

We now obtain an upper bound on the size of $\mathcal{V} \setminus \mathcal{H}$ in the following lemma whose proof is provided in Appendix H.1.

Lemma 18 *There exists a constant $C > 0$ such that when*

$$m \geq \frac{Cn\log(n/s)}{wq(1 \wedge wq)} ,$$

we have $|\mathcal{V} \setminus \mathcal{H}| \leq s$ with high probability.

This lemma implies that there exists a constant $C > 0$ such that when $m \geq \frac{Cn\log(n/s)}{wq(1\wedge wq)}$,

$$\sum_{k=1}^{K} \left|\mathcal{V}_k \setminus \hat{\mathcal{V}}_k\right| \leq |\mathcal{V} \setminus \mathcal{H}| \leq s$$

with high probability. This completes the proof of Theorem 7.

H.1. Proof of Lemma 18

We will first find the number of items that do not satisfy (H1) or (H2). Then, we will bound the number of items that do not satisfy (H3) with a given number of items that do not satisfy (H1) or (H2). From the two step proof, we can complete the proof of Lemma 18.

From Lemma 20 and the Markov inequality, it follows that with high probability, the number of items not satisfying (H1) is less than $n\exp(-C_1 np_{\max})$ for a constant $C_1 > 0$. We now bound the number of items that satisfy (H1) but not (H2). To do so, we first study properties of such items. For a given item $v \in \mathcal{V}_i$, each user's ratings can add at most one to only one entry of A_v, where the probability to add one to $A_{v'v}$ such that $v' \in \mathcal{V}_j$ is $\frac{|\mathcal{V}_j|}{m}\rho_{ij}$. Using this with Lemma 13 in (Yun and Proutiere, 2016) and Lemma 19, it is not hard to check that if item v satisfies (H1),

$$\mathbb{P}\left\{\sum_{i=0}^{K} e(\{v\},\mathcal{V}_i)\log\frac{e(\{v\},\mathcal{V}_i)/m}{\hat{p}(j,i)} \geq cnp_{\max}\right\} \leq \exp\left(-\frac{1}{2}cnp_{\max}\right) \quad \text{for all} \quad c > 0 \,. \quad (54)$$

From Lemma 2.4 in (Tsybakov, 2009), we note that

$$\sum_x (\sqrt{p(x)} - \sqrt{q(x)})^2 \leq \sum_x p(x)\log\frac{p(x)}{q(x)} , \quad (55)$$

where $\sum_x p(x) = \sum_x q(x) = 1$ and $p(x), q(x) \geq 0$ for all x. Observing $\sum_{i=0}^K e(\{v\}, \mathcal{V}_i)/m = 1$ and using (55), it follows that for $v \in \mathcal{V}_i$,

$$\begin{aligned}
&\sum_{i=0}^K e(\{v\}, \mathcal{V}_i) \log \frac{e(\{v\}, \mathcal{V}_i)/m}{\hat{p}(j,i)} + \sum_{i=0}^K e(\{v\}, \mathcal{V}_i) \log \frac{e(\{v\}, \mathcal{V}_i)/m}{\hat{p}(k,i)} \\
\geq & m \sum_{i=0}^K \left(\sqrt{\frac{e(\{v\}, \mathcal{V}_i)}{m}} - \sqrt{\hat{p}(j,i)} \right)^2 + m \sum_{i=0}^K \left(\sqrt{\frac{e(\{v\}, \mathcal{V}_i)}{m}} - \sqrt{\hat{p}(k,i)} \right)^2 \\
= & \Omega(np_{\max}) \ , \qquad (56)
\end{aligned}$$

where for the last inequality, we use the following lemma whose proof is provided in Appendix H.2.

Lemma 19 *With high probability,*

$$\left| \hat{p}(i,j) - \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right| = O\left(\frac{\log(np_{\max})^2}{\sqrt{np_{\max}}} \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right) .$$

Thus, when $v \in \mathcal{V}_i$ does not satisfy (H2),

$$\sum_{i=0}^K e(\{v\}, \mathcal{V}_i) \log \frac{e(\{v\}, \mathcal{V}_i)/m}{\hat{p}(j,i)} = \Omega(np_{\max}) \ . \qquad (57)$$

From (54), (57), and the Markov inequality, it follows that the number of items that do not satisfy (H2) is less than $n \exp(-C_2 np_{\max})$ with a constant $C_2 > 0$ with high probability.

We have shown that the number of items that do not satisfy (H1) or (H2) is less than $n \left(\exp(-C_1 np_{\max}) + \exp(-C_2 np_{\max}) \right)$ with high probability. Using this with Lemma 16 in (Yun and Proutiere, 2016) and Lemma 14, we can show

$$|\mathcal{V} \setminus \mathcal{H}| \leq 3n \left(\exp(-C_1 np_{\max}) + \exp(-C_2 np_{\max}) \right) \ ,$$

which completes the proof of Lemma 18.

H.2. Proof of Lemma 19

After the spectral partition step, we have $O(1/p)$ misclassified items. Thus,

$$\left| \mathbb{E}[\hat{p}(i,j)] - \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right| = O\left(\frac{1}{np_{\max}} \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right) . \qquad (58)$$

Using this with the Chernoff bound, it is straightforward to check that there exists a constant $C > 0$ such that

$$\mathbb{P}_1 \left\{ |\hat{p}(i,j) - \mathbb{E}[\hat{p}(i,j)]| \geq C \frac{\log(np_{\max})^2}{\sqrt{np_{\max}}} \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right\} \leq \exp\left(-2n \log(np_{\max})^2 \right) \ ,$$

which implies

$$\mathbb{P} \left\{ |\hat{p}(i,j) - \mathbb{E}[\hat{p}(i,j)]| \geq C \frac{\log(np_{\max})^2}{\sqrt{np_{\max}}} \frac{|\mathcal{V}_j|}{m} \rho_{ij} \right\} \leq \exp\left(-n \log(np_{\max})^2 \right) \ , \qquad (59)$$

where we replace $\mathbb{P}_1$ to $\mathbb{P}$ using Lemma 13. Since there are at most K^n possible configurations of clusters, we can complete the proof of Lemma 19 using the union bound, (59), and the following lemma whose proof is provided in Appendix H.3.

Lemma 20 *There exists a constant $C > 0$ such that*

$$\mathbb{P}\left\{\max_{v' \in \Gamma} \sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\} \leq \exp(-n\tilde{p}) \ .$$

H.3. Proof of Lemma 20

Pick an arbitrary $v' \in \mathcal{V}$. We first bound the tail probability of $\sum_{v \in \mathcal{V}} A_{vv'}$. Using Lemma 14, we can replace $\mathbb{P}$ with $\mathbb{P}_2$. Thus, we have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\} &\leq \mathbb{P}_2\left\{\sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\} \\
&\overset{(a)}{\leq} \max_{\lambda > 0} \frac{\mathbb{E}_2\left[\exp\left(\lambda \sum_{i=1}^n A_{vi}\right)\right]}{\exp(\lambda Cn\tilde{p})} \\
&\overset{(b)}{\leq} \max_{\lambda > 0} \frac{\exp\left(2np_{\max}(e^{\lambda} - 1)\right)}{\exp(\lambda Cn\tilde{p})} \ ,
\end{aligned}$$

where we use the Markov inequality; and $1 + x \leq e^x$ and the definition of Ψ_2 for (a) and (b), respectively. Using the Chernoff bound and the definition of $p_{\max}$, it is not hard to check that there exists a constant $c_1 > 0$ such that with high probability,

$$p_{\max} \leq c_1 \tilde{p} \ .$$

Hence, we further have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\} &\leq \max_{\lambda > 0} \frac{\exp\left(2np_{\max}(e^{\lambda} - 1)\right)}{\exp(\lambda Cn\tilde{p})} \\
&\leq \exp\left(-n\tilde{p}\left(C - 2c_1(e-1)\right)\right) \ . \qquad (60)
\end{aligned}$$

From (60) with sufficiently large $C \geq 2 + 2c_1(e-1)$, it follows that

$$\mathbb{E}\left[\left|\left\{v' \in \mathcal{V} : \sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\}\right|\right] \leq n \exp\left(-2n\tilde{p}\right) \ ,$$

which implies

$$\mathbb{P}\left\{\left|\left\{v' \in \mathcal{V} : \sum_{v \in \mathcal{V}} A_{vv'} \geq Cn\tilde{p}\right\}\right| \geq n \exp\left(-n\tilde{p}\right)\right\} \leq \exp\left(-n\tilde{p}\right) \ ,$$

where we use the Markov inequality. This completes the proof of Lemma 20.